

## XIII. Data Processing and Statistical Adjustment

### Introduction

The following guidelines detail the steps taken after the data are collected (see [Data Collection](#)). Each country's data must be processed ([coded](#), [entered](#), and [edited](#)), and then statistical adjustment ([response rate](#) calculation, missing value [imputation](#), survey [weight](#) creation, and variance estimation) can be performed. After the processing activities, the data from each country can be harmonized with those from the other countries and, after the adjustment activities, the data can be disseminated to the public as a cross-cultural dataset (see [Harmonization of Survey and Statistical Data](#) and [Dissemination of Survey and Statistical Data](#)). Substantive analysis can then be performed on the disseminated dataset.

Although the steps are the same, the flow involved in processing the survey data for paper versus computer-assisted questionnaires differs. For paper surveys, the sequential steps are as follows: code data, enter data, perform edit checks, impute missing values, create weights, build data files, and estimate [variances](#). For computer-assisted surveys, entering the data, performing edit checks, and building data files occur while the data are being collected. Then the remaining steps occur in the following order: code data, impute missing values, create weights, and estimate variances. Much burden can be eliminated with the parallel processing capabilities of computer-assisted interviewing (e.g., limited additional keying and built-in consistency checks).

Unfortunately, processing and adjustment activities often are not given adequate attention and are thus under-budgeted (e.g., editing could consume up to 40% of an entire survey budget) [\[3\]](#) [\[14\]](#). As at other stages of survey research, coders, editors, and other data processing operators may potentially produce error in the data, possibly even systematic error [\[3\]](#). Additionally, it is common for only a few errors to be responsible for the majority of changes in estimates [\[14\]](#). To lessen the effort (and possibly minimize error), checks could be performed throughout the field period (while respondent is still available) rather than waiting until the end of data collection [\[14\]](#).

These guidelines are broken down into [Data Processing Steps](#) and [Statistical Adjustment Steps](#). [Quality control](#) and documentation guidelines are applicable to both steps.

## Guidelines

### Data Processing

**Goal:** To convert the data collected during the field period into a file that can (1) be used within the organization for [quality](#) assessment of the survey implementation and (2) be made accessible to outside users for substantive research.

#### 1. Use [coding](#) to assign numeric values to survey responses.

##### *Rationale*

To statistically analyze raw responses, they must be converted into a meaningful numeric form. This process is coding. During questionnaire and instrument development, [precoding](#) should occur; that is, coding conventions and formats should be determined based on prior knowledge of the survey items (see [Survey Instrument Design](#)). Upon the collection of the data, coding decisions are revisited and possibly revised to appropriately characterize the data. Coding can be automated and/or manual. Both automated and manual coding should be evaluated at the variable, code, and coder level to detect potential error [\[3\]](#).

##### *Procedural steps*

- Decide how coding should be conducted [\[3\]](#).
  - Depending on resource and facility availability, consider centralized coding (at one location, typically the survey organization) versus decentralized coding (at several locations, typically the coders' homes);
    - Supervisory control is easier with centralized coding—which often results in higher coder reliability.
    - Centralized coding typically involves fewer coders, with each coder having a larger workload. The larger workload often results in a higher correlated [variance](#) among the coders. (For a more in-depth explanation, see a similar situation with [interviewer variance](#) and [design effect](#) in the [Interviewer Recruitment, Selection, and Training](#) chapter.)
  - Depending on the complexity of the questionnaire and variability among the response options, consider automated coding (where a computer program assigns codes) versus manual coding (where an individual assigns codes).
- Review responses and make any necessary modifications to the pre-established coding frame ([precoding](#)) in order to accurately represent the range of collected data.

- Create [nomenclatures](#) systematically. For example, the first character could represent the main coding category with subsequent characters representing subcategories [\[3\]](#).
  - With hierarchical coding structures, be especially cautious about correctly coding the first character, because errors at the higher levels are more serious.
  - If using automated coding, program the nomenclature as a dictionary database.
- Determine which variables should have standardized coding nomenclatures among countries and which could have country-specific codes.
- Create a [codebook](#) which describes how the survey responses are associated with the nomenclatures [\[3\]](#).
- From the codebook, generate a [data dictionary](#) that lists the question items (i.e., variables) with their respective response options and assigned codes. Often frequencies of each response option will be added to the data dictionary after the data has been processed.
- For automated coding, feed the responses into a computer with software that assigns appropriate codes based on matching the responses to a data dictionary [\[3\]](#).
  - Decide between using exact matching, which results in less error but also fewer assignments, or inexact matching, which has the opposite outcome.
  - Manually code any responses that are left uncoded.
- Control manual coding by using either dependent or independent verification [\[3\]](#).
  - To be effective, use coders who have equivalent coding training.
  - With dependent verification, one coder verifies and corrects another coder's work.
  - With independent verification, two coders code all responses and an additional third coder assigns codes if there is disagreement.
  - If costs allow, perform independent verification rather than dependent verification for higher [quality](#) data.
- Assess the inter-rater reliability of the coders by computing [Cohen's kappa](#) (a statistical measure that accounts for chance).
  - Kappa values between 0.7 – 0.8 are considered reliable.
  - If the kappa values are not reliable, provide additional coder training and consider revising the coding frame.
  - Consider recoding the data if the original codes are not reliable.

### ***Lessons learned***

- Although using a thorough [data dictionary](#) for automated [coding](#) generally results in less coding error, it is not always ideal to increase the detail of the dictionary, especially while the data are being processed [3]. A more descriptive data dictionary will lessen the automated coding software's ability to match and assign codes to the responses, resulting in more manual coding. For example, the data dictionary for one of the Swedish household expenditure surveys was updated 17 times while processing the data. The dictionary increased in size from 1459 to 4230 descriptions. The later versions of the data dictionary could only code up to 73% of the responses.

## **2. [Enter](#) and/or [capture](#) the data into an electronic form.**

### ***Rationale***

Like [coding](#), data entry/capture is necessary for statistical analysis. One advantage of computer-assisted questionnaires is the elimination of a separate [data entry](#) step, thus reducing the likelihood of additional processing error. When computer-assisted questionnaires are not possible, keying is often the first method of data entry that comes to mind. As technology advances, however, there are other alternatives that should be considered, such as optical character recognition, intelligent character recognition, mark character recognition, voice recognition entry, and touchtone data entry. Similarly, with developing technology, there are additional [data capture](#) possibilities, such as facsimile transmission, electronic data interchange, and e-mail transmission.

### ***Procedural steps***

- Use similar conventions in programming the data entry application as used when programming the survey instrument application (see [Survey Instrument Design](#)).
- With a paper-and-pencil questionnaire, minimize the required amount of keyer judgment by having supervisors check the responses before [data entry](#) [15]. Review the questionnaire for [16]:
  - Illegible responses.
  - Erasures.
  - Markings outside the response check box.
  - Crossed out (but still legible) responses.
  - Added response categories (None, NA, Same).
- Limit the number of individuals with access to data entry, making it easier to isolate potential problems.

- Perform independent rekey verification.
  - Have two keyers work separately and then compare their work.
  - Settle discrepancies with a computer or an adjudicator [\[3\]](#).
  - Strive to verify 100% of the data entry [\[4\]](#).
  - Look for the following keyer errors: wrong column/field and corrected/modified (misspelled) responses [\[16\]](#).
- Consider automated alternatives to key entry, including [\[3\]](#):
  - Optical character recognition (OCR) to read machine-generated characters.
  - Intelligent character recognition (ICR) to interpret handwriting.
  - Mark character recognition (MCR) to detect markings—[bubbles].
  - Voice recognition entry (VRE) to interpret oral responses.
  - Touchtone data entry (TDE) to interpret pressing numbers on the telephone keypad.
- When using automated systems [\[3\]](#):
  - Check what was captured and manually correct any errors from misreading the raw data or omitting information (e.g., with ICR).
  - Frequently recalibrate and configure scanning equipment to minimize the frequency of with which the software misreads information (e.g., with OCR).
- Do not limit [data capture](#) to traditional technology. As appropriate, reflect on the possibility of using facsimile transmission, electronic data interchange (EDI), and/or e-mail transmission [\[3\]](#).

### ***Lessons learned***

- [Data entry](#) software varies from simple spreadsheets to sophisticated applications with [quality](#) checks built-in. If the data entry software is not universal among the participating countries, then it is likely that some countries' data will be of higher quality than others'.

### **3. [Edit](#) and [clean](#) the data as a final check for errors.**

#### ***Rationale***

Editing during pre-production and data collection is a better allocation of resources than fixing errors during post-production. There can be several stages of editing [\[3\]](#). In computer-assisted surveys, the application can notify the interviewers (or respondents, if self-administered) of inconsistent or implausible responses. This gives respondents a chance to review, clarify, or correct their responses. Paper surveys can include instructions telling respondents to review their responses. Prior to [data entry/capture](#),

survey organizations can manually look for obvious errors, such as blanks. Then, during data entry/capture, editing software can be used to check for errors at both the variable and case level. Most editing takes place after data entry/capture and is described below.

### ***Procedural steps***

- Limit programming computer-assisted applications to the most important edits, so as not to increase the length of the survey or to interrupt the interview.
- Assess a random sample of each interviewer's completed paper questionnaires by examining the data entered, especially the use of skips and the frequency of missing data.
- Establish decision rules as to whether the potential errors should be accepted, changed, or flagged.
  - Flag values if they are suspicious and will need further investigation to determine if they are erroneous [3].
  - Follow-up on the suspicious values only if they could seriously affect the estimates, weighing the costs and logistics of [recontacting](#) the respondent [14].
- Using editing rules, execute a data cleaning [3]. Check for the following [3] [13] [14] [15]:
  - Unique identification number for every sample [element](#), as well as a unique identification number for each interviewer.
  - Correct number of digits for numeric variables.
  - Correct [coding](#) scheme for all variables.
  - Entirely blank variables.
  - Multi-response variables with only one response value.
  - Wild codes (e.g., out-of-range responses and unspecified response categories).
  - Internal consistency (e.g., subcategories should sum to aggregate, parents' age should be greater than children's, and males should not report pregnancies).
  - Minimum set of items filled to be considered a complete interview (including [item-missing data](#) on key variables).
  - Flow of skip patterns.
  - Reconciliation of originally collected data and [reinterview](#) data.
  - Omission or duplication of records.
- Identify fields involved in many failed edits and repair them first [5].

- Either create a code that indicates a change has been made to the collected data or keep an unedited dataset in addition to the corrected dataset [13].

### ***Lessons learned***

- [Overediting](#) may delay the release of the dataset and reduce its relevance to users [3]. Make selective editing decisions based on the importance of the sampling [element](#) or variable, the severity of the error, and the costs of further investigation. The time and money saved by implementing selective editing can be redirected to other stages of the research process.

### **Statistical Adjustment**

**Goal:** To facilitate estimates of [target population](#) attributes based on sample survey data.

4. Use [disposition codes](#) and calculate [response rates](#) based on an established, cited industry standard.

#### ***Rationale***

Response rates are one measure of survey quality and can be used to adjust survey estimates to help correct for [nonresponse](#). Therefore, reporting response rates and other outcome rates based on an established industry standard is an important part of dissemination and publication.

#### ***Procedural steps***

- Have the [coordinating center](#) provide a list of specific disposition codes and a clear description of how to code and group all sample elements during the field period (“temporary disposition codes”) and at the close of the field period (“final disposition codes”). These disposition codes will allow the standardization of rate calculations across countries.
  - Generally, disposition codes identify sample lines as “interview completed” or “non-interview.” Non-interviews are further subdivided depending upon whether the respondent is eligible or ineligible to participate in the study. Ineligible noninterviews might include the respondent being deceased, the housing unit being unoccupied, or the respondent having emigrated out of the boundaries of the study area. Eligible non-interviews include refusal to participate, noncontact, and other (defined by study).

- Disposition codes are mutually exclusive, and while each sample line may be assigned different temporary disposition codes across the field period, it will be assigned **only one** final disposition code.
- Based on an established industry standard, assign all sample [elements](#) into mutually exclusive and exhaustive categories and calculate response rates.
  - Assigning each [element](#) into predetermined final categories, those necessary in calculating a response rate, makes it possible to recalculate each country's response rate in a standard way for comparison across countries, as appropriate.
  - The World Association for Public Opinion Research/American Association for Public Opinion Research (WAPOR/AAPOR) provides one example of an established industry standard [\[1\]](#).
    - According to WAPOR/AAPOR's "Standard Definitions of Final Dispositions of Case Codes and Outcome Rates for Surveys," there are four main response rate components. These are Interviews—and three categories of Non-Interviews: Non-Interviews-Eligible, Non-Interviews-Unknown Eligibility, and Non-Interviews-Ineligible.
    - The [Tenders, Bids, and Contracts](#) appendix contains two separate templates that could be used to define the different response rate components and record counts of the different components.
    - WAPOR/AAPOR defines six separate response rates (RR1-RR6) [\[1\]](#).
      - Response rates ending in odd numbers (RR1, RR3, RR5) do not consider partially-completed interviews to be interviews. Response rates ending in even numbers (RR2, RR4, RR6) consider partially-completed interviews to be interviews.
      - RR1 and RR2 assume that all elements of unknown eligibility are eligible.
      - RR3 and RR4 estimate the percentage of elements of unknown eligibility that are actually eligible.
      - RR5 and RR6 assume that all elements of unknown eligibility are ineligible.
- Based on an established industry standard, calculate other important outcome rates such as [contact](#), [cooperation](#), or [refusal rates](#).
  - There are many different industry standards available. WAPOR/AAPOR's outcome rate calculations [\[1\]](#) are an example of one such standard.

**Lessons learned**

- Ensure that each disposition code is clearly described and reviewed during each participating country's study training. Countries may not be familiar with the specified disposition codes or the response rate terminologies. As another check, consider obtaining call records from each country early in the data collection period in order to ensure that all countries are correctly identifying different outcomes and understand the difference between temporary and final result codes. Implement all disposition codes according to project requirements.

**5. Develop [survey weights](#) for each interviewed element on the frame.****Rationale**

Depending upon the quality of the [sampling frame](#), the sample design, and patterns of [nonresponse](#), the distribution among groups of observations in a survey data set may be much different from the distribution in the population. These group differences are usually called “over representation” or “under representation.” Sampling statisticians create weights to reduce the [sampling bias](#) of the estimates and to compensate for [noncoverage](#) and nonresponse. An overall survey weight for each interviewed element typically contains three adjustments: 1) a [base weight](#) to adjust for unequal probabilities of selection ( $w_{base}$ ); 2) an adjustment for sample [nonresponse](#) ( $adj_{nr}$ ); and 3) a [poststratification](#) adjustment ( $adj_{ps}$ ) for the difference between the weighted sample distribution and population distribution on variables that are considered to be related to key outcomes. If all three adjustments are needed, the overall weight is the product of these three adjustments, or:

$$w = w_{base} * adj_{nr} * adj_{ps}$$

However, it is not always necessary to create all three weight adjustments when creating an overall survey weight. Create the adjustments only as needed. For example, if all elements had equal probabilities of selection, a base weight would not be necessary. The overall survey weight would then be the product of any nonresponse adjustment and any poststratification adjustment.

**Procedural steps**

- If necessary, calculate the base weight for each [element](#).
  - Each element's base weight is the inverse of the probability of the selection of the specified element over all stages of selection.
- If necessary, calculate the [nonresponse](#) adjustment for each element.
  - There are many ways to calculate nonresponse adjustments. This guideline will only explain one method that uses observed response rates within selected subgroups. This method is easier to calculate

- than others but assumes that all members within a specific subgroup have the same likelihood of responding. For information on other nonresponse adjustment methods, see Bethlehem [2].
- Compute response rates for mutually exclusive and exhaustive subgroups in the sample that are related to the statistic of interest.
  - The inverse of a subgroup's response rate is the nonresponse weight for each eligible, sampled element in the subgroup.
- If necessary, calculate the poststratification adjustment.
    - Multiply  $w_{base} * adj_{nr}$  to obtain a weight that adjusts for both unequal selection probabilities and sample nonresponse for each eligible element.
    - Using this weight, calculate a weighted sample distribution for certain variables related to the statistics of interest where the population distribution is known (e.g., race and sex). See [7] for a method of computing poststratification weights when the population distribution is unknown for certain subgroups ("raking" or "iterative proportional fitting").
    - Divide the known population count or proportion in each poststratum by the weighted sample count or proportion to compute  $adj_{ps}$ .
      - Example: According to 2007 estimates from Statistics South Africa, women comprise 52.2% of the total population residing in the Eastern Cape Province. Imagine the estimate of the proportion of women in the Eastern Cape from a small local survey is 54.8%. The poststratification adjustment,  $adj_{ps}$ , for female respondents in the Eastern Cape is  $.522/.548 = .953$ .
  - Multiply the needed weight adjustments together to determine an overall weight for each element on the data file.
  - Trim the weights to reduce [sampling variance](#).
    - Survey statisticians trim weights by limiting the range of the weights to specified upper and lower bounds (e.g., using no less than the 10<sup>th</sup> percentile and no more than the 90<sup>th</sup> percentile of the original weight distribution).
    - Pros/Cons: Trimming of weights produces a reduction in sampling variance but an increase in [sampling bias](#).
  - There may be weight components other than the three described in this section. Other possible weight components are country specific adjustments and weights that account for differential probability of selection for certain questionnaire sections.
  - Apply the final weight to each record when calculating the statistic of interest.

- Pros of weighting.
  - Can reduce [coverage bias](#), [nonresponse bias](#) and sampling bias at the country or study level, depending on whether the weights were designed to reflect the population of a specific country or the entire study.
- Cons of weighting.
  - Can increase sampling variance. See [Appendix C](#) for a measure of the increase in sampling variance due to weighting.

### ***Lessons learned***

- Ensure that all participating countries thoroughly document the sampling procedures and selection probabilities at every stage of selection. Countries that do not routinely employ survey weights or use [complex survey designs](#) may not be accustomed to recording and maintaining this information. Without this information, it can be very difficult to recreate base weights once data collection is complete.

## **6. Consider using [imputation](#) to compensate for [item-missing data](#).**

### ***Rationale***

Imputation is most often used to replace item-missing data and not [unit nonresponse](#). The aim is to reduce the [bias](#) in the estimate of the statistic of interest caused by item-missing data.

### ***Procedural steps***

- Select an imputation method. There are many different imputation methods available [\[8\]](#) [\[12\]](#). The methods listed below are not necessarily the best methods, but are some of the more commonly used ones.
  - Overall mean value imputation.
    - Replace the missing values for a variable with the mean value for that variable across the entire data set.
    - Pro.
      - Simple method to use.
    - Con.
      - Can distort the distribution of the variable with imputed values by creating a spike in the distribution at the mean value and [bias](#) the results.
  - Sequential hot-deck imputation.
    - Sort the data set by specific observed variables related to the statistic of interest. For example, imagine the statistic of interest

is the average yearly personal income in Spain and that it is known from previous studies that the yearly personal income in Spain is related to years of education and age. The data set would first be sorted by years of formal education and then respondent age. See if the first element on the sorted data set has a value for the variable that is to be imputed, which in the above example would be reported yearly personal income. If the first element does not have a value, impute the mean value of the variable based on the sample elements that provided the data on the statistic of interest. If the first element does have a value, keep this reported value and move to the second [element](#). The last reported value is now the “hot-deck” value. If the second element is missing a value for the specified variable, impute the “hot-deck” value. The value for the second element then becomes the “hot-deck” value for the third element, etc.

- Pros.
  - Lower cost than imputation models because no model fitting is necessary.
  - Less complex than regression imputation methods, making it more easily understood by analysts of secondary data.
  - Can reduce variance and [nonresponse bias](#).
- Con.
  - With small samples or complex missing data patterns, regression methods may produce better imputations.
- Regression imputation.
  - Create a regression model for a specific variable that predicts the value of the variable based on other observed variables in the data set. For example, one could create a regression model that predicts the number of doctor visits in the past year based on demographics such as age, sex, race, education and occupation.
  - Check that the predictor variables do not have many missing values.
  - Pro: Can provide better imputations of missing values for variables with complex missing data patterns than hot-deck methods.
  - Con: The model must be created carefully and well specified.
- Multiple imputation [11].
  - Several imputed values are created for each missing value. The imputation method must involve some randomness so that the different imputed values for a given record are not all the same.
  - When records contain different numbers of missing items, sequential regression imputation can be used. Multiple imputed data sets are created that are each based on a different trial run of a regression imputation model for each imputed item. This is an iterative process where one item is imputed using an

imputation model and then the next item is imputed with a regression model that uses the imputed values of the first item. Variation in the estimates across the trial runs allows for the estimation of both [sampling](#) and [imputation variance](#).

- Several statistical software packages are capable of multiple imputation. [IVEWare](#), a package developed at the University of Michigan and available to users without cost, is an example of one such package (<http://www.isr.umich.edu/src/smp/ive>).
- Create “Imputation Flag” fields that indicate which items for each record on the data file were imputed.

### ***Lessons learned***

- Imputation may be hard to do for all cross-cultural surveys because it takes a lot of time and decisions need to be made on what variables to impute. Imputations of variables needed for [poststratification](#) adjustments are only the minimum number of imputations necessary. Additionally, sampling statisticians advise users to avoid imputing attitudinal variables, since attitudes can easily change over time and missing data patterns can be difficult, if not impossible, to predict. Imputation models for factual variables are generally easier to specify because they are more static. Finally, it is important to check that the imputation model fits the data correctly and is well specified. A poor imputation model can actually increase the [bias](#) of the estimate, making it worse than not using imputation.
7. **When calculating the [sampling variance](#) of a [complex sample design](#), use a statistical software package with the appropriate procedures and commands to account for the complex features of the sample design.**

### ***Rationale***

The survey sample design determines the level of precision (i.e., the extent of sampling variance). Unfortunately, many statistical texts only discuss the sampling variance formulae for simple random sampling without replacement. Similarly, statistical software packages assume simple random sampling without replacement unless otherwise instructed by the user. However, compared to a simple random sample design, [stratification](#) generally decreases sampling variance while [clustering](#) increases it (see [Sample Design](#) for an in depth explanation of simple random sample, stratification, and clustering). If the correct formulas or appropriate statistical software procedures and commands are not applied, the calculation of the precision (i.e., sampling variance) of the statistic(s) of interest can be underestimated or overestimated. Therefore,

analysts are cautioned to ensure that they are applying the correct methods to calculate sampling variance, based on the sampling design.

### ***Procedural steps***

- In order to use Taylor series variance estimation, the survey data file must include, at a minimum, a final [survey weight](#), a [stratum](#) identifier, and a [sampling unit](#) identifier for each responding [sample element](#). The chosen statistical software package must have the capacity to account for survey weights, stratification, and sampling units in the estimation process [\[10\]](#).
  - If the complex sample design used clustering, the survey data should also include cluster identifiers for each responding sample element.
  - In order to estimate the sampling variance within a stratum, at least two selections must be made within the stratum. For a sampling design that selects only one [primary sampling unit](#) (PSU) per stratum, the sampling variance cannot be estimated without [bias](#). In “one PSU per stratum” designs, the PSUs are arranged after data collection into a set of [sampling error computational units](#) (SECUs) that can be grouped into pairs for purposes of estimating approximate variances. If a participating country uses a sample design that selects only one PSU per stratum, the survey data must include the SECU of each element to make variance estimation possible.
- When a survey data file is supplied with a series of [replicate](#) weights plus the final survey weight, balanced repeated replication or jackknife repeated replication can be used to estimate variances (see [Appendix B](#)).
- When calculating estimated means and variances using statistical software (e.g., SAS, STATA), use the appropriate procedures and commands to account for the complex sample data. For example, SAS version 9.1.3 features the `surveyfreq` and `surveymeans` procedures with `strata` and `cluster` commands to account for complex designs.

### ***Lessons learned***

- Not all countries/cultures may have access to statistical software packages (e.g., STATA, SAS, SPSS); therefore, it may be necessary to arrange for reduced fees or for centralized analysis.

## Data Processing and Statistical Adjustment

### 8. Implement [quality control](#) checks at each stage of the data processing and statistical adjustment processes.

#### *Rationale*

Ensuring [quality](#) is a vital part of each stage of the survey lifecycle. Even after data collection is complete, the survey organization must continue to implement quality control measures to help reduce or eliminate any errors that could arise during the post-production procedures discussed above. If the emphasis on quality is relaxed during these latter activities, all of the time and money spent on maintaining quality during the previous stages of the survey lifecycle will be compromised.

#### *Procedural steps*

- Continually evaluate processing activities, such as the number of responses that were [coded](#) automatically; were changed after dictionary updates; and were coded in error due to coding mode, category, or dictionary updates [\[3\]](#).
- Use [data entry](#) programs to perform keying [quality](#) control checks. Have human analysts check for representativeness and outliers [\[15\]](#).
- Monitor [editing](#), possibly using some of Granquist's and colleagues' (1997) key process statistics [\[3\]](#). Some of their process statistics are as follows (where objects can refer to fields, characters, or records):
  - Edit failure rate = # of objects with edit failures / # of objects edited (estimate of amount of verification).
  - Recontact rate = # of recontacts / # of objects edited (estimate of number of recontacts).
  - Correction rate = # of objects corrected / # of objects edited (estimate of the effect of the editing process).
- Remove any identifying information from the production data. For example, remove any names and addresses attached to each responding unit.
- Check to see if missing data can be filled from other respondent information. For example, the sampling frame may contain information, such as age or race/ethnicity, that the respondent refused to provide.
- When possible, use auxiliary data (e.g., census or population files) for post-survey adjustments and to enhance the precision of the survey

estimates. For example, population files could be used to create [nonresponse](#) weighting adjustment categories.

- Compare the sum of the [base weights](#) of the initially sampled elements to the count  $N$  of units on the sampling frame. If the sample was selected with probabilities proportional to size, then the sum of base weights is an estimate of  $N$ . If an equal probability sample was selected within strata or overall, then the sum of sample weights should be exactly equal to  $N$ .
- Assign a second sampling statistician to check the post-survey adjustment methodology and the statistical software syntax of the survey's primary sampling statistician.

### ***Lessons learned***

- Make certain that all identifying information is removed from the dataset before making it publicly available. In some surveys, this may require that detailed geographic identifiers be removed. One of the authors of these guidelines was a study participant in a survey that publicly released a data set that included variables which made it easy to personally identify each respondent. Thankfully, her identity was not stolen, but the principles of the Helsinki Declaration (see [Ethical Considerations](#)) and the ethical requirements of social science human subjects research were egregiously violated.

## **9. Document the steps taken in data processing and statistical adjustment.**

### ***Rationale***

Over the course of many years, various researchers will analyze the same survey dataset. In order to provide these different users with a clear sense of how and why the data were collected, it is critical that all properties of the dataset be documented.

Documentation will help secondary data users better understand post-survey statistical adjustments that can become quite complex, such as [imputation](#) and [sample weighting](#). A better understanding of these adjustments will help ensure that secondary data users correctly interpret the data. In addition, post-survey documentation will indicate whether the survey organization that conducted the survey met benchmarks agreed to in the contract by the [coordinating center](#) and the survey organization.

### ***Procedural steps***

- Document the procedures and [quality](#) indicators of the data processing. Examples include:
  - Training protocol for data [coding](#) and [entry](#) staff.
  - Who performed the coding and entered the data.
  - Staff evaluation protocol for data coding and entry.
  - What items were coded or re-coded.
  - Measure of coding reliability (i.e., [Cohen's kappa](#)).
  - Data entry verification protocol.
  - Data entry accuracy rate.
  - Any [cleaning](#) of open-ended responses (e.g., to remove identifying information, correct typographical errors, standardize language).
  - Open-ended responses cleaning protocol.
  - How the raw data were corrected during the cleaning process.
  
- Describe how the sample identification numbers/codes were assigned to each [sampling unit](#).
  - For internal documentation, describe how the unique sample identification number/code was assigned for internal use data sets (e.g., 0600500200101: first 2 digits identify the country, the next 3 digits identify the area segment, the next 3 digits identify the sample [replicate](#), the next 3 digits identify the household, the final two digits indicate the respondent where 01=main respondent selected and 02=second respondent selected).
  - For internal and external documentation, describe how a different unique sample identification number/code was assigned for public use data sets. This public use sample identification number is used to prevent disclosing a respondent's identity (see [Ethical Considerations](#)).
  
- If values were imputed for the study, clearly describe the imputation method that was used.
  - Specify the imputation flags that indicate whether a value was imputed for a particular variable.
  
- If weights were generated for the study, clearly explain how each individual weight adjustment was developed.
  - Each explanation should include both a written description and the formula used to calculate the weighting adjustment. Below are examples of the first sentence of an explanation for different weight adjustments. These are not meant to be exhaustive explanations, and the documentation of each adjustment should include further written descriptions and formulas.
    - The [base weight](#) accounted for oversampling in the Wallonia region (Belgium) strata.

- The [nonresponse](#) adjustment was the inverse of [response rate](#) in each of three regions – Vlanders, Wallonia and Brussels.
- The [poststratification](#) adjustment factor adjusts weighted survey counts to totals from Denmark's population register by sex, education and age.
- As of March 1, 2004, a random half of the outstanding [elements](#) in the field were retained for additional follow-up efforts, and this subsample of elements was given an extra weight adjustment factor of  $W=1/.5=2.0$ .
  - If additional weight adjustments were used to calculate a final weight, provide a clear description of how these components were created. Examples of additional weight components are country specific adjustments and adjustments that account for differential probability of selection for certain questionnaire sections.
  - Address whether there was any trimming of the weights and, if so, the process used to trim the weights.
  - Address whether a procedure was used for scaling of the weights (e.g., population (N), population (N in 1000s), sample size (centered)).
  - If a replicated weighting method was used (i.e., Jackknife Repeated Replication or Balanced Repeated Replication – see [Appendix B](#) for more), provide the replicate weights for variance estimation.
  - Clearly describe when each of the survey weights and adjustments that were developed for the study should be used in data analysis.
- For [complex survey sample data](#), identify the [cluster](#) and [stratum](#) assignment variables made available for sampling error calculations. For instance:
  - The variable that identifies the stratum to which each sample ID belongs.
  - The variable that identifies the sampling cluster to which each sample ID belongs.
    - If the sample design has multiple stages of selection, document the variables that identify each unique [sample element's](#) primary [sampling unit](#) (PSU), [secondary sampling unit](#) (SSU), etc.
    - If Balanced Repeated Replication variance estimation was used, identify the stratum-specific half sample variable, i.e., a field that identifies whether a unit is in the [sampling error computation unit](#) (SECU) 1 or 2.
- If the risk of disclosing respondent identities is low, consider providing the different weight components on public use data sets.
- Discuss whether the survey met the requirements (e.g., [response rates](#), number of interviews) outlined in the contract.

- If the requirements were not met, provide possible reasons why the survey failed to meet these requirements.
- Provide ways to improve the quality of future surveys of a similar design.

### ***Lessons learned***

- Secondary users of survey data often have a hard time understanding when and if they should use weights in their analyses. This issue is exacerbated in many cross-cultural surveys, where participating countries may apply different nonresponse and postratification adjustment strategies. Without a clear documentation of how the each country created their survey weights and when to use each of the weights in data analysis, the chance of secondary users either not applying or incorrectly applying weights and producing estimates that do not accurately reflect the respective [target population](#) greatly increases. The chance of a secondary user incorrectly applying weights increases in a cross-cultural survey, where participating countries may apply different nonresponse and postratification adjustments. Therefore, clear and understandable documentation of the post production processes is extremely important.

## Appendix A

### First-stage Ratio Adjustments

- If using a multistage, [probability proportional to size](#) design, calculate the ratio adjustments for each stage prior to the final stage of selection.
  - This adjustment stabilizes the estimates over different selections of [primary sampling units](#) (PSUs) in the [stratum](#) so that the weighted total is consistent over realizations of the sample design.

$$\text{Estimated Stratum Population Total} = \frac{\text{Population Total in Selected PSU}}{\text{Probability of Selecting PSU}}$$

$$w_{ratio\_adj} = \frac{\text{Stratum Population Total from Frame}}{\text{Estimated Stratum Population Total}}$$

## Appendix B

### Estimating ratio means or other complex statistics when the sample size is not fixed

- Whenever the sampling size is not fixed, use the Taylor-Series estimation or one of the replicated methods (Balanced Repeated Replication (BRR) or Jackknife Repeated Replication (JRR) to estimate ratio means or other complex statistics.
  - Taylor Series estimation – computes the [sampling variance](#) of an approximation to a complex function like a ratio or regression coefficient. (See [\[9\]](#) for the exact formulas.)
    - Pro: used by most statistical software packages.
    - Cons:
      - Requires analytic manipulations and computation of derivatives (but these have been done by developers of the software packages for common type of estimates).
      - Not useful if estimate cannot be expressed as a function of sample totals.
      - Taylor-Series estimates in most software packages do not account for the variability of [nonresponse](#) adjustments.
  - Balanced Repeated Replication (or Half-Sample Replication).
    - This method assumes a paired selection design (i.e., 2 PSUs per stratum) and selects  $H^*$  half sample [replicates](#) ( $H^*$  is the smallest multiple of 4 greater than or equal to the number of strata) by deleting one [PSU](#) from each stratum according to the pattern in a [Hadamard matrix](#). Each remaining element in the half sample receives a replicate weight of two. Fay's method of BRR [\[6\]](#) is an alternative that retains both PSUs in a pair but modifies their [survey weights](#).
  - Pros:
    - More useful for complex estimates such as medians than Taylor-Series.
    - Easily applied to user-specified statistics like differences or ratios of domain means.
    - Accounts for variability due to multiple steps in adjustment more easily than does Taylor-Series.
  - Cons:
    - Best used only with a paired selection stratification design.
    - Appending replicate weights to each record increases file size.
    - Combining of strata and PSUs is sometimes done to reduce number of replicates. This must be done carefully to avoid [biased](#) variance estimates.
- Jackknife Repeated Replication.

- This method creates a replicate by dropping a PSU from one stratum and weights up the other PSUs in the stratum to maintain the sampling distribution across the strata.
- Pros.
  - More useful for complex estimates than the Taylor series.
  - Easily applied to user-specified statistics like differences or ratios of domain means.
  - Can handle designs other than paired selection.
  - Accounts for variability due to multiple steps in adjustment more easily than does Taylor-Series.
- Cons.
  - Not appropriate for the variance of quantiles like the median.
  - Appending replicate weights to each record increases file size.
  - Combining of strata and PSUs is sometimes done to reduce number of replicates. This must be done carefully to avoid biased sampling variance estimates.

## Appendix C

### Loss in Precision of Estimate Due to Weighting in Household Surveys

- While overall [survey weights](#) help decrease three different sources of [bias](#) ([coverage](#), [nonresponse](#) and [sampling](#)), the variability of the weights also can increase the [sampling variance](#) in household surveys. The formula below is a simple model to measure the loss in precision ( $L_w$ ) due to weighting. It assumes that the weights and the variable of interest are not related.

$$\blacksquare \quad L_w = \left[ \frac{\sum_{i=1}^n w_i^2}{\left( \sum_{i=1}^n w_i \right)^2} \right] (n) - 1$$

- For example, if  $L_w = .156$ , then the sampling variance of the estimate increased by 15.6% due to differential weighting.
- $L_w$  can also be calculated for subgroups.
- N.B. This formula does not apply to surveys of institutions or business establishments where differential weighting can be efficient.

## Glossary

<b>Base weight</b>	The inverse of the probability of selection.
<b>Bias</b>	A systematic difference between the survey estimate of the population parameter and the true value in the population.
<b>Clustering</b>	A sample design where the <a href="#">elements</a> of the <a href="#">sampling frame</a> are grouped into clusters. The clusters are then sampled and data is collected from one or more elements within each sampled cluster.
<b>Codebook</b>	A document that provides question-level metadata that are matched to variables in a dataset. Metadata include the elements of a data dictionary, as well as basic study documentation, question text, universes (the characteristics of respondents who were asked the question), the number of respondents who answered the question, and response frequencies or statistics.
<b>Coding</b>	Translating nonnumeric data into numeric fields.
<b>Cohen's kappa</b>	A statistical measure that accounts for chance.
<b>Complex survey data (or designs)</b>	Survey data sets (or designs) based on stratified single or multistage samples with <a href="#">survey weights</a> designed to compensate for unequal probabilities of selection or <a href="#">nonresponse</a> .
<b>Contact rate</b>	The proportion of all <a href="#">elements</a> in which some responsible member of the housing unit was reached by the survey.
<b>Cooperation rate</b>	The proportion of all <a href="#">elements</a> interviewed of all eligible <a href="#">units</a> ever contacted.
<b>Coordinating center</b>	A research center that facilitates and organizes cross-national research activities.
<b>Coverage</b>	The proportion of the <a href="#">target population</a> that is accounted for on the <a href="#">sampling frame</a> .
<b>Coverage bias</b>	The systematic difference between the expected value (over all conceptual trials) of a statistic and the <a href="#">target population</a> value because some <a href="#">elements</a> in the target population do not appear on the <a href="#">sampling frame</a> .

<b>Data capture</b>	Data collection.
<b>Data cleaning</b>	Identifying and correcting errors (defined by <a href="#">editing</a> rules) in the dataset.
<b>Data dictionary</b>	Question or variable-level metadata, including variable names, labels, and data types.
<b>Data entry</b>	The process of transferring verbal or written responses to an electronic form, for use by a computer.
<b>Design effect</b>	The impact of the <a href="#">complex survey design</a> on <a href="#">sampling variance</a> measured as the ratio of the sampling variance under the complex design to the sampling variance computed as a simple random sample.
<b>Disposition code</b>	A code that indicates the result of a specific call attempt or the outcome assigned to a sample <a href="#">element</a> at the end of data collection (e.g., noncontact, refusal, ineligible, complete interview).
<b>Editing</b>	Altering data recorded by the interviewer or respondent to improve the <a href="#">quality</a> of the data (e.g., checking consistency, correcting mistakes, following up on suspicious values, deleting duplicates, etc.). Sometimes this term also includes <a href="#">coding</a> and <a href="#">imputation</a> , the placement of a number into a field where data were missing.
<b>Element</b>	A single unit of the sampling frame.
<b>Hadamard matrix</b>	Square arrays of + and – that define balanced half samples. Such matrices exist for any multiple of four. Pluses [+] mean keep the first <a href="#">PSU</a> and minuses [-] keep the second PSU in the stratum. Therefore, the first half sample identified in the matrix below keeps the first PSU in strata 1, 2, 3 and the second PSU in stratum 4.

<b>Hadamard matrix for 4 half samples</b>	Half Sample	Stratum			
		1	2	3	4
	1	+	+	+	-
	2	+	-	-	-
	3	-	-	+	-
4	-	+	-	-	

<b>Imputation</b>	Computation methods that, using some protocol, assign a proxy value for each <a href="#">missing data item</a> .
<b>Imputation variance</b>	That component of overall variability in survey statistics that can be accounted for by imputation.
<b>Interviewer variance</b>	That component of overall variability in survey statistics that can be accounted for by the interviewers.
<b>Item-missing data</b>	The absence of information on individual data items for a sample <a href="#">element</a> successfully measured on other items.
<b>Nomenclatures</b>	Set of code numbers.
<b>Nonresponse</b>	The failure to obtain measurement on sampled <a href="#">units</a> .
<b>Nonresponse bias</b>	The systematic difference between the expected value (over all conceptual trials) of a statistic and the <a href="#">target population</a> value due to differences between respondents and nonrespondents on that statistic of interest.
<b>Overediting</b>	Extensive <a href="#">editing</a> that becomes too financially costly for the amount of error that is being reduced
<b>Precoding</b>	When designing the questionnaire and survey instrument, determine <a href="#">coding</a> conventions and formats of survey items (especially the close-ended questions) based on existing coding frames or prior knowledge of the <a href="#">survey population</a> .
<b>Primary sampling unit (PSU)</b>	A <a href="#">unit</a> sampled at the first stage of selection.
<b>Probability proportional to size</b>	A sampling method that changes “the first- and second-stage selection chances in such a way that when multiplied together the probability is equal for every <a href="#">element</a> , and the sample size is the same from one sample to the next.”
<b>Poststratification</b>	A statistical adjustment that assures that sample estimates of totals or percentages (e.g. the estimate of the percentage of men in living in Mexico based on the sample) equal population totals or percentages (e.g. the estimate of the percentage of men living in Mexico based on Census data). The adjustment cells for poststratification are formed in a similar way as strata in sample selection,

but variables can be used that were not on the original [sampling frame](#) at the time of selection.

<b>Quality</b>	Achieving excellence for all components related to the data.
<b>Quality control</b>	Process focused on fulfilling quality requirements.
<b>Recontact</b>	Having another staff member (often a supervisor) attempt to speak with the respondent after the interview is reported, in order to verify that the interview was completed according to the specified protocol.
<b>Refusal rate</b>	The proportion of all elements in which a housing unit or respondent refuses to do an interview or breaks off interviews of all potentially eligible elements.
<b>Reinterview</b>	The process or action of interviewing the same respondent twice to assess reliability (simple response variance).
<b>Replicates</b>	Systematic probability subsamples of the full sample.
<b>Response rate</b>	The number of complete interviews with reporting units divided by the number of eligible reporting units in the sample.
<b>Sampling bias</b>	The systematic difference between the expected value (over all conceptual trials) of an unweighted sample estimate and the <a href="#">target population</a> value because some <a href="#">elements</a> on the <a href="#">sampling frame</a> have a higher chance of selection than other elements.
<b>Sample element</b>	A selected <a href="#">unit</a> of the <a href="#">target population</a> that may be eligible or ineligible.
<b>Sampling error computational units (SECUs)</b>	<a href="#">PSUs</a> in 'one PSU per stratum' sampling designs that are grouped in pairs, after data collection, for purposes of estimating approximate <a href="#">sampling variances</a> .
<b>Sampling frames</b>	Lists or procedures intended to identify and allow access to all elements of a <a href="#">target population</a> .
<b>Sampling unit</b>	<a href="#">Elements</a> or clusters of elements considered for selection in some stage of sampling. For a sample with only one stage of selection, the sampling units are the same as the elements. In multi-stage samples (e.g., enumeration areas,

then households within selected enumeration areas, and finally adults within selected households), different sampling units exist, while only the last is an element. The term [primary sampling units](#) (PSUs) refers to the sampling units chosen in the first stage of selection. The term [secondary sampling units](#) (SSUs) refers to sampling units within the PSUs that are chosen in the second stage of selection.

<b>Sampling variance</b>	A measure of how much a statistic varies around its mean over all conceptual trials as a result of the sample design only. This measure does not account for other sources of variable error such as <a href="#">coverage</a> and <a href="#">nonresponse</a> .
<b>Secondary Sampling Unit (SSU)</b>	A unit sampled at the second stage of selection.
<b>Stratification</b>	A sample design that divides each the sampling frame into mutually exclusive and exhaustive strata and places each element on the frame into one of the strata. Independent selections are then made from each strata, one by one, to ensure representation of subgroups of the <a href="#">survey population</a> in the sample.
<b>Stratum</b>	A mutually exclusive group of <a href="#">elements</a> on a <a href="#">sampling frame</a> .
<b>Survey population</b>	The actual population from which the survey data are collected, given the restrictions from data collection operations.
<b>Survey weight</b>	A statistical adjustment created to compensate for <a href="#">complex survey designs</a> with features including, but not limited to, unequal likelihoods of selection, differences in <a href="#">response rates</a> across key subgroups, and deviations from distributions on critical variables found in the <a href="#">target population</a> from external sources, such as a national Census.
<b>Target population</b>	The finite population for which the survey sponsor wants to make inferences using the sample statistics.
<b>Unit nonresponse</b>	A sample <a href="#">element</a> that has little or no information because the individual declined the invitation to participate in the survey. Also known as a nonrespondent.

**Variance**

A measure of how much a statistic varies around its mean over all conceptual trials.

## References

- [1] AAPOR Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, Version 4. Retrieved January 9, 2008, from [http://www.aapor.org/uploads/Standard\\_Definitions\\_04\\_08\\_Final.pdf](http://www.aapor.org/uploads/Standard_Definitions_04_08_Final.pdf)
- [2] Bethlehem, J. G. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. Groves, D. Dillman, J. Eltinge, & R. Little, (Eds.) *Survey Nonresponse*, (chap. 18). New York: Wiley.
- [3] Biemer, P., & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.
- [4] Federal Committee on Statistical Methodology. (1983). *Statistical policy working paper 9: Contracting for surveys*. Washington, DC: Office of Management and Budget.
- [5] Fellegi, I. P., & Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71(353), 17-35.
- [6] Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6(3), 223-239.
- [7] Kalton G. (1983). *Compensating for missing survey data*. University of Michigan, Survey Research Center, Institute for Social Research.
- [8] Kalton, G., & Kasprzyk, D. (1986). Treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- [9] Kish, L. (1965). *Survey sampling*. New York: Wiley & Sons.
- [10] Lepkowski, J., & Bowles, J. (1996). Sampling error software for personal computers. *The Survey Statistician*, 35, 10-17.
- [11] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2<sup>nd</sup> edition, New York: John Wiley.
- [12] Marker, D. A, Judkins, D. R., & Winglee, M. (2001). Large-scale imputation for complex surveys. In Groves, R., Dillman, D., Eltinge, J., & Little, R. (Eds.), *Survey nonresponse* (chap. 22). New York: Wiley, 2001.

- [13] Office of Management and Budget. (2006). *Standards and guidelines for statistical surveys*. Washington, DC: Office of Information and Regulatory Affairs, OMB. Retrieved June 9, 2008, from [http://www.whitehouse.gov/omb/infoereg/statpolicy/standards\\_stat\\_surveys.pdf](http://www.whitehouse.gov/omb/infoereg/statpolicy/standards_stat_surveys.pdf)
- [14] Statistics Canada. (2003). *Statistics Canada quality guidelines*. Montreal: Statistics Canada. Retrieved June 9, 2008, from <http://www.statcan.ca/english/freepub/12-539-XIE/index.htm>
- [15] United Nations. (2005). *Household surveys in developing and transition countries*. NY: United Nations, Department of Economic and Social Affairs.
- [16] Wurdeman, K. (1993). Quality of data keying for major operations of the 1990 census. Unpublished manuscript.

## Further Reading

- Groves, R. M. (1989). *Survey errors and survey costs*. Hoboken, NJ: Wiley & Sons.
- Groves, R. M., Dillman, D. A., Eltinge, J. L. & Little, R. J. A. (Eds.), (2002). *Survey nonresponse*. New York: Wiley.
- Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: Wiley & Sons.
- Horvitz, D. G., & Thompson, D..J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-665.
- Kish, L., & Hess, I. (1959). On variances of ratios and their differences in multistage samples. *Journal of the American Statistical Association*, 54, 416-446.
- Lessler, J., & Kalsbeek, W. (1992). *Nonsampling error in surveys*. New York: Wiley.
- Lyberg, L., Biemer, P., Collins, M., de Leeuw, E., Dippo, C., Schwarz, N., et al. (Eds.). (1997). *Survey measurement and process quality*. New York: Wiley.
- Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85-95.
- Smith, T. W. (2003). A review of methods to estimate the status of cases with unknown eligibility. Report of the Standard Definitions Committee for the American Association for Public Opinion Research.
- Worcester, R., Lagos, M., & Basanez, M. (2000). Problems and progress in cross-national studies: Lessons learned the hard way. Paper presented at the WAPOR/AAPOR annual conference, Portland, OR.