Supporting Document A
Organization of Metadata

Version 1.4

Issued: 09 Mar 06

Census Bureau Standard

Definitions for Survey and Census Metadata

Authored for:

Cynthia Z.F. Clark
Associate Director
for Methodology and Standards

USCENSUSBUREAU

Helping You Make Informed Decisions

Document Management & Control [1]

Version	Issue Date	Approval	Description
1.0	19 Dec 02	M&S Council	Initial Release
1.1	13 Feb 03	M&S Council	Reissue
1.2	09 Jul 03	Associate Directors	Initial Concurrence
1.3	27 Dec 04	Configuration Mgr.	Reformatted to comply with Census Bureau Identity Standard and Quality Program Document Management Plan
1.4	09 Mar 06	Configuration Mgr.	Inserted hyperlink for main standard.

[1] **The most current version of this document is maintained on the Census Bureau Intranet and may be accessed from the Quality Management Repository.**

Organization of Metadata[2]

1. Planning

1.1 Abstract – Survey Name/Designation, General Objectives/Purpose, and Information.

Survey – activity that collects or acquires data from census, sample survey, administrative records, derived statistical activity, etc.[1]

Objectives – purposes for which information is required, stated within the context of the program, research problem or hypotheses that gave rise to the need for information.[1]

Uses – decisions to be made based on collected information and how information will support decisions.[1]

Users – organizations, agencies, and groups expected to use the information.[1]

Concepts – subjects of inquiry and analysis of interest to users. General characteristics or attributes of a statistical unit or of a population of like statistical units.[1]

1.2 Contact

1.2.1 **Survey Sponsor and Legal Authority**: agency(s) or organizations responsible for sponsoring survey, survey design, collection, processing, and dissemination under U.S. Code or contract.

1.2.2 **Sponsor and Bureau Persons**: sponsor and Census Bureau contact persons with names, telephone numbers, and email addresses, etc.

1.3 Major Variables of Interest/Data Items in Publication

Key variables – main classification variables (e.g., geography, demographic attributes, economic attributes, industry etc.) of units to be studied.[1]

Characteristics of interest – main variables having different possible values for different units of analysis.[2]

[2] Sources of the definitions are indicated by superscripts and listed on the last page of this attachment.

1.4 Survey Description

1.4.1 **Budget**: Funding and costs of project with estimates allocated for components of project/survey, including fixed overheads and those that are sample size related.

Cost function – a mathematical expression showing the cost of conducting a survey in terms of the sample sizes and unit costs.[1]

1.4.2 Statistical Precision and Accuracy

1.4.2.1 Concepts of Statistical Precision and Accuracy

Estimation Error – for a given estimator, the difference between the value of the estimate and the true population value being estimated. Includes both sampling error and nonsampling error.

Sampling Error – for a given estimator, the difference between an estimate based on a sample and the estimate that would result if the sample were to include the entire population.

Nonsampling Error – for a given estimator, the difference between the estimate that would result if the sample were to include the entire population and the true population value being estimated.

Expected Value/Mean Value (of an estimator) – the average of the estimator over all possible samples of the same design.

1.4.2.2 Measures of Statistical Precision and Accuracy

Bias – the difference between the expected value of an estimator and the true population value being estimated.

Sampling Bias – for a given estimator, the difference between the expected value of the estimator based on a sample and the estimate that would result if the sample were to include the entire population.

> **Example**: an area frame with unintentional overlaps, due perhaps to ill-defined boundaries, may cause overcoverage when only a sample of segments are listed and interviewed. When the entire list is included, these overlaps are easily discovered and eliminated.

Sampling Variance (or sampling error variance) – the expected squared deviation of an estimator from its mean value, with the expectation taken over all possible samples of the same design.

Standard Error – the square root of the estimated variance.

Coefficient of Variation – the standard error divided by the mean value of the estimator.

Mean Squared Error – the expected squared deviation of an estimator from the true population value being estimated, equal to the variance plus the squared bias (a statistical measure of accuracy that reflects both sampling and nonsampling errors).

1.4.3 **Survey Schedule/Collection Period**: Periodicity of the survey, the collection period date(s), ongoing survey reference period, and data release date(s) etc.

1.4.3.1 **Reference Period and Release of Results** (timeliness)

Timeliness – length of time between data availability and the event or phenomenon it describes (context of value and use).[1]

1.4.3.2 **Recall and Recall Period**

1.4.4 **Geographic Scope**: National and subnational level areas covered by the survey, such as region, state, metropolitan area, county, city, (codes) etc., for which estimates will be made.

1.4.5 **Collection Mode and Type of Respondent/Unit of Observation**

Computer Assisted Telephone Interviewing (CATI) – Method or mode of data collection using telephone interviews in which the questions (to be asked) are displayed on a computer screen and responses are entered directly into the computer.

Computer Assisted Personal Interviewing (CAPI) – Method or mode of data collection consisting of the interviewer asking questions displayed on a portable computer screen and entering the answers directly into the computer.

Collection Mode – Mail out/mail in, CATI, CAPI, Personal Visit (PV), etc.

Type of Respondent – Housing units, persons (self/proxy), establishments, etc.

1.4.6 **Quality Control** (including goals)

Statistical Quality Control – refers to the statistical procedures used to control the quality of the output of some production process, including the output of data. The review and evaluation of the production process also includes inspection of specifications, data collection procedures, review of sample output, processing, tabulation of data products, etc.[4]

Reinterview – repeated measurement of the same unit designed to detect and deter falsification.

1.4.7 **Data Collection Training and Nonresponse Requirements**

1.4.7.1 **Training**: Planned training for field representatives to collect, code, and transmit data. Includes planning of manuals for both field representatives and trainers.

1.4.7.2 **Nonresponse**: Plan for following up nonrespondents.

1.4.8 **Background and Special Features**: Includes historical background of survey and special features such as redesign, data collection mode, budget increases/decreases, major initiatives/revisions over time, etc. Identifies statistics currently available and suspect data needing improvements.

1.4.9 **Policy Considerations/Standards**

1.4.9.1 **Confidentiality**

Confidentiality – protection/entrustment of respondent's information.[1]

Disclosure – situation in which an individual may use published statistics to identify either an individual or business that has provided information under a pledge of confidentiality.[12]

Disclosure Limitation – process for protecting the confidentiality of information that has been provided by respondents.[12]

1.4.9.2 **Questionnaire Testing**

1.4.9.3 **National or International Standards** (e.g., NAICS, SOC, FIPS, etc.)

1.4.9.4 **Uses of Estimates Extrapolated from Census Bureau Tables**: Data users who create their own estimates using data from Census Bureau tables should cite the Census Bureau as the source of the original data only.[12]

1.4.10 Other Quality Issues

1.4.10.1 **Relevance** – qualitative assessment of the value contributed by the data. Value is characterized by the degree to which the data serve to address the purposes for which they are produced and sought by users (including mandate of the agency, legislated requirements, etc.)[1]

1.4.10.2 **Accessibility** – availability of information from the holdings of the agency, also taking into account the suitability of the form in which the information is available, the media of dissemination, the availability of metadata, and whether the user has reasonable opportunity to know it is available and how to access it (also includes affordability).[1]

1.4.10.3 **Interpretability** – user ease in understanding and properly using and analyzing the data or information. Adequacy of definitions of concepts, target populations, variables and terminology underlying the data, and information on any limitations of the data.[1] Tabulation of data clearly defined. Definitions current and changes over time conveyed.

1.4.10.4 **Coherence** – degree to which data from a statistical program, and data from other data sets or statistical programs, are logically connected and complete (consistent over time, across products and programs). Also concepts are logically distinguishable from similar, but not identical, concepts of other statistical programs.[1] Shortcomings of current statistics regarding data needs identified.

1.5 System

1.5.1 System Description

1.6 Project

1.6.1 Project Description

2. Content

2.1 **Target Population** (Population of Interest) – the clearly defined set of elements, or units of analysis, which are the individual members about which information is wanted and estimates are required.[1] The target population is also referred to as the population or universe of interest.

2.2 **Sampled Population** – the population to be sampled. It is usually the same as the target population, but, for reasons of practicability or convenience, may be different than the target population.[11] These differences, if any, should be explained.

2.3 **Variables of Interest** – the main characteristics associated with each of the elements of the population of interest.

2.4 **Data Sets and Products Description/Specs**

2.4.1 **Survey Data Sets** (Content)

2.4.1.1 **Dataset Description** – covers

Microdata Record Layout

Macrodata/Table Outlines

Geographic Level

2.4.2 **Products**

2.4.2.1 **Product Description** – covers

Public Use File

Press Releases

Reports

Tables

2.5 **OMB Clearance and Expiration Dates** – All federal censuses and surveys require Office of Management and Budget clearance approval numbers and expiration dates.

3. **Design**

3.1 **Sampling Frame**

3.1.1 **Frame Concepts**

Sampling Frame – any list or device that, for purposes of sampling, de-limits, identifies, and allows access to the sampling units, which contain elements of the sampled population.[1] The frame may be a listing of persons, housing units, businesses, records, land segments, etc.[2] One sampling frame or a combination of frames may be used to cover the entire sampled population.

List or List Frame – a collection of sampling units that have been numbered or otherwise identified.

Administrative Records – data collected for the purpose of carrying out various programs (e.g., tax collection) used in the creation and maintenance of sampling frames.[1]

Master Address File (MAF) – is the Census Bureau's permanent list of addresses for individual living quarters nationwide.

3.1.1.1 Name

3.1.1.2 Description

3.1.1.3 Creation Date

3.1.2 Frame Update Specifications

3.1.2.1 **Births** – Additions to the sampling frame that did not exist or were not identified when the sampling frame was initially constructed.

3.1.2.2 **Deaths** – Deletions to the sampling frame that are no longer believed to be part of the sampled population.

3.1.2.3 **Mergers and Splits** – Results of sampling frame units merging together or separating apart over time.

3.1.3 Alternative Sources for Target Population and Evaluation

3.2 Frame Coverage of Sampled Population

Coverage – the extent to which a frame includes all the elements of the sampled population.[1]

3.2.1 **Overcoverage** – The extent to which a frame includes more elements than the sampled population; including duplicate elements.

3.2.2 **Undercoverage** – The extent to which a frame includes fewer elements than the sampled population.

3.3 Sample Design

Sample Design – the sampling plan and estimation procedures. The sampling plan is the methodology used for selecting the sample. The estimation procedures are the algorithms or formulas used for obtaining estimates of population values of interest from the sample data and for estimating the reliability of these estimates.[9]

Sampling – the selection of a set of sampling units from a sampling frame. The set of sampling units selected is referred to as the sample. If all the units are selected, the sample is referred to as a census.

Probability Sample – sample selected such that every sampling unit on the sampling frame has a known, nonzero probability of being included in the sample. In probability sampling, the reliability of the resulting population estimates can be evaluated.[9]

3.3.1 **Sample Concepts/Description (Includes objectives/limitations)** – General text description of sample design including stratification, stages of sampling, clustering, sample size and allocation, method of selection, rotation sampling scheme used, date of sample, subsampling nonrespondents, etc.

3.3.2 Sample Specifications/Documentation

3.3.2.1 Multi-Stages of Sampling

Multi-Stage Sampling – a sample of clusters is selected and then a subsample of units is selected within each sample cluster. If the subsample of units is the last stage of sample selection, it is called a two-stage design. If the subsample is also a cluster from which units are again selected, it is called a three-stage design, etc.[2]

Primary Sampling Units – selected first units that are clusters of reporting units from which there is subsampling to obtain reporting units in a multistage sample.[1]

3.3.2.2 Stratification Used

Stratifying Variables – variables whose joint values are used to classify a sampling frame into several classes (called strata) from each one of which a sample is drawn independently. Both numerical and nonnumerical variables may be used for creating strata. Variables from outside sources, such as administrative records, censuses, etc., may be used as stratifying variables.[4]

Stratification – dividing the sampling frames into subsets (called strata) before the selection of a sample within each of the subsets for statistical efficiency, for production of estimates by stratum, or for convenience. Stratification is done such that each stratum contains units that are relatively homogeneous with respect to variables that are believed to be highly correlated with the information requested in the survey.[1]

3.3.2.3 Clustering

List Sample – selection of sample from total list of units (clustered or not) in which the sampling units have been numbered or otherwise identified.[2]

Cluster Sample – sampling in which the units of analysis are considered as grouped into clusters, and a sample of clusters is selected. The selected clusters determine the units to be included in the sample, and the sample may include all units in the selected cluster or a subsample in each selected cluster.[2]

Area Sample – sampling units are individual land areas (segments) which can be identified on a map. Boundaries of each segment must be clearly defined so they can be identified by enumerators in the field.[2]

3.3.2.4 Sample Sizes and Allocation

Sample Size – size of sample determined in relation to the required precision and available budget for observing the selected units.[1]

Allocation – the method in determining how the sample should be distributed. In stratified sampling, it usually refers to the determination of the units selected from each stratum. In cluster sampling, it refers to the decision as to the number of clusters to be selected and the size of the sample in each cluster.[2]

Proportional Allocation – allocation in which the ratio of the number of sampled sampling units to the total number of sampling units is the same for each stratum.[11]

Optimum Allocation – in stratified sampling with a linear cost function, allocation in which the variance of the estimated mean or total is minimized for a specified cost, or the cost is minimized for a specified variance of the estimated mean or total. For a given stratum, the sample size is proportional to the product of the stratum's total number of sampling units and standard deviation, and it is inversely proportional to the square root of the cost per unit for the stratum.[11]

Neyman Allocation – a special case of optimum allocation in which the variance of an estimated mean or total is minimized for a specified total sample size. For a given stratum, the amount of the total sample size that is allocated depends on the relative size of the product of the stratum's total number of sampling units and standard deviation.[11]

3.3.2.5 Methods of Sample Selection (including subsampling)

Judgment Sampling – the sample elements are carefully selected to provide a "representative" sample, but selection bias could arise with "expert" choice.[4]

Simple Random Sampling – for a sample of size n, each of the possible combinations of n sampling units that may be formed from a population of N sampling units has the same chance of selection as every other combination of n units. Also every sampling unit will have the same chance of selection as every other sampling unit.[2]

Probability Proportional to Size (PPS) Sampling – a method of sample selection in which the units are selected with unequal probability; the probability for each unit being proportionate to its measure of size. The measure of size for a unit is a number assigned to that unit in advance of selection, which is believed to be highly correlated with the variables to be collected in the survey.[2]

Stratified Sampling – the method of sampling from a sampling frame which has been stratified. At least one sampling unit must be selected from each stratum. Probabilities of selection can be different from stratum to stratum.[2]

Stratified Random Sampling – a special case of stratified sampling in which a simple random sample is taken in each stratum.[11]

Systematic Sampling – a method of sample selection in which the sampling frame is listed in some order and every k th element is selected for the sample, beginning from a random start between 1 and k.[2]

Sampling with Replacement – a method of sample selection in which a sample is obtained by first selecting one sampling unit from the sampling frame, replacing it, then making a second selection and replacing it before making a third selection, etc., until n selections have been made. A particular unit could be included more than once in the sample and possibly up to n times.[2]

Sampling without Replacement – a method of sample selection in which a sample is obtained by selecting one sampling unit from the sampling frame and, without replacing it, selecting one of the remaining sampling units; then continuing this process until n different selections have been made. A unit can be included only once in any sample.[2]

Self-weighting Sample – a sample in which every sampling unit on the sampling frame has the same chance of selection, although unequal probabilities may have been used at various stages of sampling.[2]

Double Sampling – a method of sample selection in which a sample is obtained by selecting a large sample in the first phase and then a subsample of the first-phase sample in the second phase. Information needed for sample design or estimation is collected from the large first-phase sample and then used in the design of the second-phase sample or in the final estimation.[4]

Subsampling Nonrespondents – an economical method for reducing nonresponse bias in which new attempts are made to obtain responses from a subsample of sampling units that did not provide responses to the first attempt.[10]

3.3.2.6 **Preparation and Maintenance of Sample**: Procedures for preparing and maintaining the list of selected sampling units for data collection. These include providing additions for births and designating deletions for deaths, as well as noting any other changes over time.

3.4 Listing Operations Specification

3.5 Observation Models

3.6 Estimation Specifications

Estimation – process that consists of approximating unknown population parameters by using information from a data set. Estimated parameters include constants such as totals, means, ratios, regression coefficients and variances.[1]

3.6.1 Point Estimates and Interval Estimates Requirements/Specs

Estimator – a function that is used to calculate an estimate of an unknown population value from the results of a given sample.[13]

Estimate – a numerical quantity calculated from sample data and intended to provide information about an unknown population value.[2]

Expected Value – see definition in 1.4.2.1.

Unbiased Estimator – an estimator having the property that its average over all possible samples of the same design is equal to the true value (i.e., the expected value is equal to the true value).[2]

Bias – see definition in 1.4.2.2.

Confidence Interval – a range about a given estimator that has a specified probability of containing the result of a complete enumeration.[10]

3.6.1.1 Imputation Requirements/Specs

Imputation – process used to resolve problems of missing, invalid, or inconsistent responses identified during editing. Responses or missing values on the edited record are changed to ensure that a plausible, internally coherent record is created.[1]

Item Nonresponse – occurs when a respondent provides some, but not all, of the requested information, or if the reported information is not useable.[14]

Partial Nonresponse – A partial interview is when some but not all items have responses. A partial interview is treated as a "unit response" when a sufficiently accurate response is obtained for only some of the data items required from a respondent and meets some minimum threshold level. A partial interview is treated as a "unit nonresponse" when this threshold is not met.[1]

Unit Nonresponse – when the sampled unit response does not meet a minimum threshold and is classified as not having responded at all;[1] failure to make measurements or obtain observations on a listing unit selected for inclusion in a sample.[4]

3.6.1.2 **Weighting (Basic and Multi-Stage) Specs**

Sampling Weight – weight assigned to a given sampling unit that is equal to the inverse of the unit's probability of being included in the sample, which is determined by the sample design. This weight may include a factor due to subsampling.[1]

3.6.1.3. **Nonresponse Adjustment Specs**

Unit Nonresponse – see definition in 3.6.1.1.

Unit Nonresponse Adjustment – an adjustment to responding units' sampling weights due to nonresponding units to improve the accuracy of the estimate.[1]

3.6.1.4 **Modeling/Estimation Techniques Specs** (multiplicative, additive, subdomain, etc.)

Auxiliary/Independent Information – information from other sources than the survey itself (e.g., administrative records, census projections, etc.) used as supplementation in the estimation stage.[1, 2]

Ratio Estimation – a method of estimating from sample data, using the ratio x/y where both x and y are estimated totals based on sample data, or a variation of this formula. One variation is to produce an estimate of a population total, X, by using the formula xY/y, where x and y are estimated totals and Y is an independently known population total.[2]

Consistent Estimator – an estimator in which the probability that it is in error by more than any given amount tends to zero as the sample becomes large. A common example of such an estimator is the ratio estimator.[11]

Small Area Estimators – special estimation methods that "borrow strength" from related areas (or domains) to minimize the mean square error of the resulting estimator.[1]

Post-stratification – Stratification of selected sampling units based on data collected after sampling, rather than before sampling.

3.6.1.5 **Control Total Specs**: Documented methodology for developing and maintaining this independent auxiliary information used for improving estimation.

3.6.1.6 **Composite Estimation Specs**

Composite Estimation – an estimation method that exploits the correlation over time in periodic surveys with a large sample overlap between occasions. It treats the data from previous occasions as auxiliary variables.[1]

3.6.1.7 **Seasonal Adjustment Specs**

Seasonal Adjustment – consists of estimating seasonal factors and applying them to a time series to remove the seasonal variations in the estimates. The variations represent the composite effect of climatic and institutional factors that repeat with certain regularity within the year.[1]

3.6.2 **Variance Estimation Requirements/Specs**

Sampling Error – see definition in 1.4.2.1.

Standard Error – see definition in 1.4.2.1.

Coefficient of Variation – see definition in 1.4.2.2.

Sampling Variance – see definition in 1.4.2.2.

Relative Variance (Rel-variance) – the square of the coefficient of variation.[2]

Finite Population Correction Factor – the term in the formula for the variance of a simple random sample which reflects the effect of the proportion of the population in the sample.[2]

Replication – method of variance estimation based on the variation between estimates of the population value of interest that are produced from each of several subsamples (replicates) of the parent sample. The three methods – random groups, balanced half samples, and jackknife – differ only in the way the replicates are formed.[6]

Generalized Variance Function – method of variance estimation which models an estimator's variance as a function of the estimator's expectation.[6]

Taylor Series Approximation – method of variance estimation to obtain an estimator of variance for certain nonlinear estimators (e.g., the classical ratio estimator) through linearization of the nonlinear estimator via Taylor Series approximation.[6]

Tables/Model Parameters/Design Effects – presentation of variance estimates.

3.6.3 Nonsampling Error Requirements/Specs

Nonsampling Error – see definition in 1.4.2.1.

Bias – see definition in 1.4.2.2.

Mean Squared Error – see definition in 1.4.2.2.

3.6.3.1 Coverage Error

Coverage Error – error due to omissions, erroneous inclusions, and duplications of units in the frame used to conduct the survey; also, for household surveys, any omissions or duplicates within the households.[1]

3.6.3.2 Nonresponse Error

Nonresponse Error – error caused by survey failure to get a response to one or possibly all of the questions. Indirect measures include the detail disposition rates (unweighted and weighted) of all the selected sample cases during data collection. Direct measures may require nonresponse follow-up.[1]

3.6.3.3 Measurement Error

Measurement Error – error when the response received differs from the "true" value due to the respondent, the interviewer, the questionnaire, the mode of collection, or the respondent's record-keeping system(s).[1, 5]

Re-interview – repeated measurement of the same unit to estimate measurement/response error.[5]

3.6.3.4 **Processing Error**

Processing Error – error during data editing, coding, capture (keying and scanning), imputation, and tabulation.[1]

3.7 **Questionnaire Specification** (Name, Concepts and Definitions)

Questionnaire – a set of questions designed to collect information from a respondent. A questionnaire may be interviewer-administered or respondent-completed, using paper-and-pencil methods for data collection or computer-assisted modes of completion.[1]

3.7.1 **Contact Specification and Information Sources (Respondent or Target Source) and Disposition of Nonrespondents**: Procedures for contacting respondents, criteria for determining acceptable respondents or records, and criteria for determining disposition of nonrespondents.

3.7.2 **Measurement Instrument** (questions and script)

3.7.2.1 **Questionnaire Design** (text, order of questions, response choices, skip patterns, recall period, etc.)

Type of Questionnaire (paper, electronic)

3.7.2.2 **Questionnaire Development, Testing, and Update** (lab/cognitive/usability testing and field testing)

Reports of Testing Design Options

Update Specifications (changes)

OMB Clearance and Expiration Date for Questionnaire testing (generic): See definition in 2.4.

3.7.2.3 **Data Collection Mode** (mail, CATI, CAPI, fax, etc.)

Case Selection Specs (assignment)

Case Preparation Specs

3.8 **Planned Observation Register Documentation** (admin)

3.9 **Data Preparation/Processing and Edit Specs**

3.9.1 **Keying and Scanning Specs**

3.9.2 **Edit Specs**

Editing – application of checks that identify missing, invalid, or inconsistent entries or that point to data records that are potentially in error. Some of these checks involve logical relationships that follow directly from the concepts and definitions, such as ratio edits, planned and tested. Others are more empirical in nature or are obtained as a result of the application of statistical tests or procedures.[1]

3.9.3 **Coding Specs**

Coding – conversion of data collection information into an electronic format suitable for use by subsequent processes.[1]

3.10 **Quality Control Specs** (design, data collection mode, and processing)

3.10.1 **Production Standard**

3.11 **Reinterview Design/Methodology/Specs**

Reinterview – see definition in 1.4.6.

3.11.1 **Reinterview Instrument**

3.11.2 **OMB Clearance and Expiration Date for Reinterview Survey** – see definition in 2.4.

3.12 **Instrument Deployment**

4. **Data Collection Procedures and Reports/Registers** (Production and Field Tests)

Data Collection – any process whose purpose is to acquire or assist in the acquisition of data. Collection is achieved by requesting and obtaining pertinent data from individuals or organizations via an appropriate vehicle.[1]

4.1 **Questionnaire** (includes option to post automated questionnaire to the Internet with its skip patterns), i.e., production instrument – see definition in 3.7.

4.2. **Data Collection Description** (includes length of interview, periodicity, respondent rules, respondent sampling, dependent interviewing, bounding techniques, enhancements over time, etc.)

4.2.1 **Procedures** (Data Collection Manual)

4.2.2 **Field Representative Training Procedures** (Manual)

Field Representative Experiences and Expertise

4.2.3 **Outreach and Promotion Procedures**

4.2.4 **Listing Operation Procedures**

4.2.5 **Mailout/Check-In/Check-Out Procedures**

4.2.6 **Coverage Procedures: Coverage** – see definition in 3.2.

4.2.7 **Noninterview and Nonresponse Follow-up Procedures**

Unit Nonresponse – see definition in 3.6.1.1.

Nonresponse Follow-up – when operational and cost constraints permit, nonrespondents can be followedup (as a complete enumeration or on a subsample basis). This procedure increases the response rate and can help ascertain to some extent whether respondents and nonrespondents are similar in the characteristics measured.[1]

4.2.7.1 **Partial Nonresponse Procedures**

Partial Nonresponse – see definition in 3.6.1.1.

4.2.8 **Data Capture Mode Description**

4.2.9 **Measurement Error Procedures**

Measurement Error – see definition in 3.6.3.3.

4.2.10 **Quality Control (QC) Data Collection Procedures**

4.2.11 **Reinterview Procedures**

Reinterview – see definition in 1.4.6.

4.2.12 Survey Preparation Procedures

4.2.13 Field Observation Procedures

4.3 Reports and Records

4.3.1 Interviewer Notes

4.3.2 Data Collection and QC Reports

4.3.2.1 **Data Collection Reports** (includes disposition counts of contacts, respondents, nonrespondents, etc.)

4.3.2.2 QC Reports

Quality Control – see definition in 1.4.6.

4.3.3 Reinterview Reports

Reinterview – see definition in 1.4.6.

4.3.4 Editing Reports

Editing – see definition in 3.9.1.

4.3.5 Coding Reports

Coding – see definition in 3.9.2.

4.3.6 Planned Observation Register

4.3.7 Cost Reports

4.3.8 Other Regional Office and Processing Center Organization Reports

4.4 Production of Final Anomalous (unusual) Observation Register

4.4.1 Treatment of Overcoverage/Undercoverage

Overcoverage/Undercoverage – see definitions in 3.2.

4.4.2 Treatment of Nonresponse

Unit Nonresponse – see definition in 3.6.1.1.

Nonresponse Follow-up – see definition in 4.2.7.

4.4.2.1 Treatment of Partial Nonresponse

Partial Nonresponse – see definition in 3.6.1.1.

4.4.3 Counting of Anomalous Observations

4.4.4 Derived Variables

4.5 Data Transmission

4.5.1 Systems

4.5.2 Mode

4.5.3 Problems

5. Data Processing (Computing Systems) Production and Field Tests

5.1 Data Processing Description

5.2 Data Preparation and Editing Processing

5.2.1 Data Capture Processing

5.2.1.1 Forms Check In

5.2.1.2 Data Entry (includes keying/imaging/OCR)

5.2.2 Editing Processing

Editing – see definition in 3.9.2.

5.2.2.1 PreEdits

5.2.2.2 Edits

5.2.3 Coding Processing

Coding – see definition in 3.9.3.

5.2.4 Survey Preparation Processing

5.2.5 Quality Assurance Methods

5.3 Imputation Processing

Imputation – see definition in 3.6.1.1.

5.3.1 Partial Nonresponse

Partial Nonresponse – see definition in 3.6.1.1.

5.3.2 Full Nonresponse

Unit Nonresponse – see definition in 3.6.1.1.

5.4 Coverage Processing

Coverage – see definition in 3.2.

5.5 **Nonresponse and Nonresponse Follow-up Processing**

Nonresponse – see definition in 3.6.1.1.

Nonresponse Follow-up – see definition in 4.2.7.

5.6 **Outreach and Promotion Processing**

5.7 **Estimation/Tabulation Processing** (including reweighting)

5.7.1 **Dictionary of Variable Definitions** – Definition of variables by collected data elements from questionnaire for generation of files and tables

5.7.2 **Estimation/Tabulation**

Estimation – see definition in 3.6.

Unbiased Estimator – see definition in 3.6.1.

Bias – see definition in 1.4.2.

5.7.3 **Variance Estimation Processing**

5.8 **Measurement Error Processing**

Measurement Error – see definition in 3.6.3.3.

5.9 **Quality Control Processing**

QC – see definition in 1.4.6.

5.9.1 **Data Collection**

5.9.2 **Processing** (Keying, etc.)

5.10 **Reinterview Processing**

Reinterview – see definition in 1.4.6.

5.11 **Interviewer Administrative Processing**

5.11.1 **Cost and Progress Processing**

5.11.2 **Payroll Processing**

6. **Data Quality, Analysis, and Evaluation**

Quality – the elements of quality consist of the relevance, accuracy, timeliness, accessibility, interpretability, and coherence of the data.[1]

Data Analysis – process of transforming raw data into useable information, often presented in the form of a published analytical article, in order to add value to the statistical output.[1]

Data Quality Evaluation – process of evaluating the final product in light of the original objectives of the statistical activity, in terms of the data's accuracy or reliability. Two general types of data quality evaluations are certification/validation and sources of error studies.[1]

Sources of Error Studies – provide quantitative information on specific sources of error in the data.[1]

6.1 Estimated Measures of Precision

Standard Error – see definition in 1.4.2.2.

Coefficient of Variation – see definition in 1.4.2.2.

Sampling Variance – see definition in 1.4.2.2.

Rel-Variance – see definition in 3.6.2.

Replication – see definition in 3.6.2.

Generalized Variance Function – see definition in 3.6.2.

Taylor Series Approximation – see definition in 3.6.2.

6.2 **Nonsampling Errors and Analysis** – see definition in 3.6.3.

Bias – see definition in 1.4.2.2.

6.2.1 Coverage Error

Coverage Error – see definition in 3.6.3.1; indicators include error rates due to: births, deaths, out of scope, unclassifieds, misclassifieds, duplications, etc.[7]

6.2.2 **Nonresponse Rates and Nonresponse Error** (including Partial [Item] Nonresponse and Nonresponse Follow-up)

Nonresponse Error – see definition in 3.6.3.2.

6.2.3 Measurement/Response Error

Measurement Error – see definition in 3.6.3.3; indicators include error rates due to: edit failure, interviewer error, instrument error, collection mode effect, record check studies, etc.[7]

6.2.4 Processing Error

Processing Error – see definition in 3.6.3.4; indicators include error rates due to: keying, coding, edit failures, imputation, reclassification, etc.[7]

6.2.5 **Overall QC Procedures Results and Analysis**

Quality Control – see definition in 1.4.6.

6.2.6 **Reinterview Results and Analysis**

Reinterview – see definition in 1.4.6.

.

6.2.7 **Additional Sources of Error and Analysis**

6.3 **Analysis of Estimation Components**

Estimation Errors – errors which may be introduced due to the use of estimators that introduce biases, deliberately or otherwise, e.g., some small area estimators.[1]

6.3.1 **Weighting** (Basic and Multi-Stage)

Sampling Weight – see definition in 3.6.1.2.

6.3.1.1 **Nonresponse Adjustment** – see definition in 3.6.1.3.

6.3.2 **Modeling/Estimation Techniques** (multiplication, additive, subdomain, etc.) – see definition in 3.6.1.4.

6.3.3 **Control Totals** – see definition in 3.6.1.5.

6.3.3.1 **Sources**

6.3.3.2 **Derivation Methodology**

6.3.3.3 **Impact on Sources of Error**

6.3.4 **Composite Estimation** – see definition in 3.6.1.6.

6.3.5 **Seasonal Adjustment** – see definition in 3.6.1.7.

6.4 **Disclosure Analysis and Research**

Confidentiality – see definition in 1.4.9.1.

Disclosure – see definition in 1.4.9.1.

Disclosure Limitation – see definition in 1.4.9.1.

6.4.1 **Macrodata Analysis**

Cell Suppression – disclosure limitation technique where sensitive cells are generally deleted from a table and flags are inserted to indicate this condition.[1]

6.4.2 Microdata Analysis

Swapping – disclosure limitation technique which involves selecting a sample of records, finding a match in the data base on a set of predetermined variables, and swapping all other variables.[8]

Recoding – disclosure limitation technique which involves collapsing/regrouping detail categories of a variable so that the resulting categories are safe.[1]

Topcoding – disclosure limitation technique which involves limiting the maximum value of a variable allowed on the file.[1]

6.5 Other Analysis

6.5.1 Macrodata Analysis

6.5.2 Microdata Analysis

6.6 Cost Analysis

6.7 Data Quality/Evaluation Reports

Documentation – constitutes a record of the statistical activity, including the underlying concepts, definitions, and methods used in the production of the data. It also includes descriptions of influences affecting comparability of data and of data quality.[1]

6.7.1 **Relevance** – see definition in 1.4.10.1.

6.7.2 **Accuracy and Precision** – see definitions in 1.4.2; final report use of standard errors including statement on level of significance for comparisons.

6.7.3 **Timeliness** – see definition in 1.4.3.1.

6.7.4 **Accessibility** – see definition in 1.4.10.2.

6.7.5 **Interpretability** – see definition in 1.4.10.3.

6.7.6 **Coherence** – see definition in 1.4.10.4.

7. **Data Dissemination** – release to users of information obtained through statistical activity.[1]

7.1 External Dissemination

7.1.1 System

7.1.2 Mode

7.1.3 Problems

7.1.4 **Usability Testing for Data Dissemination**

7.2 **Product**

Name, Series, Version, Release Date, etc.

7.3 **Archiving**

7.4 **Internal Dissemination**

7.5 **Production of Final Observation Register**

8. **Data Products**

8.1 **Released Data**

8.2 **Active Data**

8.3 **Archived Data**

8.4 **Analytical Reports**

REFERENCE TO DEFINITION SOURCE:

[1] *Statistics Canada Quality Guidelines*, Third Edition, 1998

[2] *Sampling Lectures*, Department of Commerce, Bureau of the Census, 1968

[3] *Sample Survey Methods and Theory, Vol I*, Hansen, Hurwitz, and Madow, 1965

[4] *Encyclopedia of Statistical Sciences*, Kotz, Johnson, and Read, 1988

[5] *Survey Errors and Survey Costs*, Groves, 1989

[6] *Introduction to Variance Estimation*, Wolter, 1985

[7] *Quality of Establishment Surveys*, *OMB Statistical Policy Working Paper #15*, July 1988

[8] *Statistical Disclosure Limitation Methodology*, *OMB Statistical Policy Working Paper #22*, May 1994

[9] *Sampling of Populations: Methods and Applications*, Levy and Lemeshow, 1991

[10] *Survey Sampling*, Kish, 1965

[11] *Sampling Techniques,* Cochran, 1977

[12] *Census Bureau Standards for Statistical Reliability Statements for Census and Survey Data Tables*, Census Bureau Methodology and Standards Council, 2001

[13] *Merriam* Webster's Collegiate Dictionary, 1996

[14] *Measuring and Reporting Sources of Error in Surveys, OMB Statistical Policy Working Paper #31*, July 2001