Draft BTS Statistical Standards and Guidelines

September 5, 2022

Contents

Chapter 1 – Introduction

The Bureau of Transportation Statistics (BTS), like other federal statistical agencies, establishes official standards to guide the methods and procedures for the collection, processing, data quality assurance, storage, and presentation of statistical data. Standards and guidelines define the professional basis and the level of quality and effort expected in all statistical activities, including those of contractors. The standards ensure consistency and data quality among studies conducted by the Department of Transportation (DOT) and provide users clear documentation of the methods and principles employed in the development, collection, processing, analysis, and dissemination of statistical information.

Quality of data has many faces. Primarily, it must be *useful* to its users. Usefulness (relevance) is achieved through a series of steps starting with a planning process that links user needs to data requirements. It continues through acquisition of data that are *objective* and *accurate* in measuring what it was designed to measure and produced in a *timely* manner. Finally, the data must be made *accessible* and *easy to interpret* for the users. In a more global sense, data programs also need to be *complete* and *comparable* (to both other data systems and to earlier versions). The creation of data that address all facets of quality requires effort through all the development phases from the initial data program objectives, through program design, collection, processing, and dissemination to the users. This document is intended to help management and data program "owners" achieve data quality through that sequential process.

1.1 Legislative Requirements

Among other responsibilities, the Director of BTS “shall issue guidelines for the collection of information by the Department [of Transportation] that the Director determines necessary to develop transportation statistics and carry out modeling, economic assessment, and program assessment activities to ensure that such information is accurate, reliable, relevant, uniform, and in a form that permits systematic analysis by the Department.” (49 USC § 6202(b)(3)(B)(viii))

BTS guidelines must conform with guidance from the Office of Management and Budget (OMB), whose Director shall “coordinate the activities of the Federal statistical system to “ensure … the integrity, objectivity, impartiality, utility, and confidentiality of information collected for statistical purposes.” (44 USC § 3504(e)) OMB legislation includes the Paperwork Reduction Act, the Information Quality Act, and the Foundations for Evidence-Based Policymaking Act (Evidence Act).

1.2 Department of Transportation Guidance

The Secretary of Transportation issued a policy statement on information quality and information dissemination quality guidelines in October 2019 (DOT-OST-2019-0135 at https://www.transportation.gov/dot-information-dissemination-quality-guidelines). This statement directs people with questions concerning “statistical disseminated information [to] contact the Bureau of Transportation Statistics.”

1.3 FCSM Data Quality Framework

To assess and maximize the quality of its statistics, BTS adopts the *Framework for Data Quality* published by the Federal Committee on Statistical Policy (FCSM). The FCSM, charted by OMB in 1975, "serves as a resource for OMB and the federal statistical system to inform decision making on matters of statistical policy and to provide technical assistance and guidance on statistical and methodological issues." (https://www.fcsm.gov/about/) The FCSM Data Quality Framework "provides a common foundation upon which federal agencies can make decisions about the management of data products throughout their lifecycle by identifying and mitigating key data quality threats, evaluating trade-offs among different quality dimensions where necessary, applying accepted methods at an appropriate level of rigor, and accounting for and reporting on the quality of data products and outputs. These activities all support appropriate and effective use of data." (https://www.fcsm.gov/assets/files/docs/FCSM.20.04_A_Framework_for_Data_Quality.pdf) Further information about the Framework and its use are published on the FCSM website (https://www.fcsm.gov/resources/data-quality-subcommittee/).

The framework identifies three domains of data quality:

1. The first domain is utility which refers to the usefulness of the data to the intended users' needs.
2. The second is objectivity which refers to whether the data are accurate, reliable, and unbiased and presented in a clear, unbiased, and accurate way.
3. The third is integrity. Integrity refers to protection of information from manipulation or unauthorized access and keeping to rigorous scientific standards.

Within each domain, FCSM defined dimensions. Those dimensions are shown and defined in Table 1.1 from the FCSM Data Quality Framework report.

Table 1.1 FCSM Dimensions of Data Quality. Table taken from the FCSM Data Quality Framework

Domain	Dimension	Definition
Utility	Relevance	Relevance refers to whether the data product is targeted to meet current and prospective user needs.
	Accessibility	Accessibility relates to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable to data users.
	Timeliness	Timeliness is the length of time between the event or phenomenon the data describe and their availability.
	Punctuality	Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release.
	Granularity	Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any of a number of characteristics (*e.g.* (demographic, socio-economic).
Objectivity	Accuracy and reliability	Accuracy measures the closeness of an estimate from a data product to its true value. A related concept is reliability, which characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.
	Coherence	Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data.
Integrity	Scientific integrity	Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence.
	Credibility	Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.
	Computer and physical security	Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.
	Confidentiality	Confidentiality refers to a quality or condition of information as an obligation not to disclose that information to an unauthorized party.

These domains align with the OMB Guidelines for Ensuring Information Quality and are addressed throughout this document. It is important to consider these domains and dimensions about the management of data products throughout their lifecycle.

1.4 Applicability of the Standards

The standards and guidelines in this *Manual* apply to DOT data collections or surveys whose purposes include the description, estimation, or analysis of the characteristics of groups. This includes the development, implementation, or maintenance of methods, technical or administrative procedures, or information resources that support those purposes. Certain standards and guidelines also apply to the compilation of data from external sources and to the dissemination of DOT information products.

These guidelines apply to all statistical information that is disseminated by agencies of the Department of Transportation (DOT) to the public using the "dissemination" definition in the OMB guidelines. That definition exempts several classes of information from these guidelines. Major types of exempted information are listed below. A more detailed list is provided in section 4 of the DOT Information Dissemination Quality Guidelines, of which this document is a subsection.

- Information distributed to a limited group of government employees, agency contractors, or grantees or intended to be limited to intra- or inter- agency use.
- Archival records that are inherently not "active."
- Public filings.
- Contents of the National Transportation Library that are not products of DOT-funded research or DOT-funded data collections.
- Materials that are part of an adjudicatory process.
- Hyperlinked information.
- Opinion offered by DOT staff when made clear it is opinion and not fact or the Department's views.
- Press releases and other information of an ephemeral nature, advising the public of an event or activity of a finite duration.
- Procedural, operational, policy, and internal manuals prepared for the management and operations of Dot that are not primarily intended for public dissemination.

DOT disseminated data and data products contain a lot of information provided by "third party sources" like the states, industry organizations, and other federal agencies. These guidelines apply to that disseminated data and data products unless exempted for other reasons discussed above. However, DOT guidelines indicating design, collection, and processing methods do not apply to data acquisition steps performed by non-federal sources. Steps performed by federal sources outside DOT before providing the data to DOT will be governed by the agency's own guidelines. For data provided to DOT by third party sources, these guidelines primarily emphasize disseminating information about data quality, the DOT processing methods, and analysis of the data provided to the users.

1.5 Overview of the Statistical Guidelines

These statistical guidelines have been built from the BTS Statistical Standards Manual and Guide to Good Statistical Practice in the Transportation Field. Updates to good statistical practice since the publication of these two documents have also been incorporated.

OMB defines quality in terms of utility (usefulness of information to intended users), objectivity in presentation and in substance, and integrity (protection of information from unauthorized access or revision). BTS addresses the OMB quality criteria in the following fashion:

• Utility – the planning standards and guidelines (Chapter 2) stress user involvement, while the dissemination standards and guidelines (Chapter 5) emphasize accessibility and transparency to users.

• Objectivity – objectivity in substance, through sound statistical methods, is the focus of the standards and guidelines. Chapter 5 deals with objectivity in presentation.

• Integrity – the standards and guidelines (Chapters 2 through 5) incorporate compliance with existing DOT policies for maintaining data security and protecting confidentiality.

Finally, Chapter 6 contains standards and guidelines to ensure the quality of DOT statistical processes and products by monitoring compliance with these standards.

Overall, the guidelines are ordered to match the process of building a data program. Chapter 2 covers the planning of data programs, followed by Chapter 3 which discusses the collection of data. Next, data processing is covered in Chapter 4. Ultimately, the goal is to disseminate information, and data dissemination is covered in Chapter 5. Finally, Chapter 6 covers ongoing efforts to maintain and evaluate data quality.

REFERENCES

1. 49 U.S.C. Chapter 63.
2. Paperwork Reduction Act, 1995
3. Information Quality Act. Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001. Pub. L. 106–554 (Dec.. 21, 2000).
4. Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. 115-435 (Jan. 14, 2019)
5. Secretary of Transportation, policy statement on information quality and information dissemination quality guidelines, DOT-OST-2019-0135, October 2019, https://www.transportation.gov/dot-information-dissemination-quality-guidelines
6. Federal Committee on Statistical Methodology. 2020. *A Framework for Data Quality.* https://www.fcsm.gov/assets/files/docs/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Chapter 2 - Planning Data Programs

Data collection programs must be designed to meet both internal and external user needs and the agency's legislative mandates. This chapter covers the planning and design of data collection programs, including:

- Establishing data needs and data collection program objectives (Section 2.1),
- Defining data requirements (Section 2.2)
- Identifying existing sources of data (Section 2.3)
- Designing a data collection plan (Section 2.4)

2.1 Data Program Objectives

Defining clear, specific program objectives, and identifying data users and data user needs will help guide the program development and ensure relevance. The evidence act defines relevant statistical information as 'processes, activities, and other such matters likely to be useful to policymakers and public and private sector data users" (Pub. L. No. 115-435, 132 Stat. 5529, 2018 codified at 44 USC 3563(d)(4)). Just as user needs change over time, the objectives of the data program will need to change over time to meet new requirements.

Important definitions:

Data program – any collection of information that is used as a source by any Government entity to disseminate information to the public, along with the planning, collection, processing, and evaluation of that collection.

Program owner – the organizational entity whose strategic plan and budget will guide the creation and continued maintenance of the data program.

Users – people or organizations who use the data and data products.

Objectives – describe what federal programs and external users will accomplish with the information.

GUIDELINES

Guideline 2.1.1: Definition of Data Needs and Objectives. Identify the intended audience of the final data products and establish program objectives in clear, specific terms that address data user needs and data analysis goals before initiating data program development. Write data program objectives in terms of the questions that need to be answered by the data; not in terms of the data itself.

Whenever possible, consult with data users to identify data user needs (see Guideline 2.1.2). Every data program objective should be traceable to user (including federal) needs and support the mission of the department. Two examples are (1) NHTSA's Fatality Analysis Reporting System (FARS) developed and used to track overall highway safety trends and to evaluate the effectiveness of highway safety improvements efforts; (2) BTS' Commodity Flow Survey (CFS) developed and used to provide timely, accurate, and credible information on the movement of goods.

The definition of data needs should include:

- What data items are needed and how they will be used,
- The precision level required for estimates,
- The format, level of detail, and types of tabulations and outputs, and

- When and how frequently users need the data.

Guideline 2.1.2: Consultation with Data Users and Providers. To help ensure relevance, the program owner should develop and update the data program objectives in partnership with critical users, stakeholders, and data providers. The owner should have a process to regularly update the program as user needs change.

- OMB requires publication of a Federal Register notice requesting public comments for all proposed information collections, administered by a federal agency, that would collect data from ten or more persons outside the federal government within a year,
- Expand consultations with data users and providers to include other means for collecting comments and suggestions, such as individual meetings, focus groups, presentations at conferences and workshops, cognitive testing, and pretests/pilot tests.
- Develop a standard procedure for reaching out to stakeholders for comments and feedback
- When revising an established data collection program, review any previous evaluation studies for information relating user needs to current program performance. Ensure that any changes to the data program maintains data continuity to allow users to monitor trends over time.

Guideline 2.1.3: Choice of How to Meet Data Requirements. Review related studies and data collection programs before beginning detailed planning for the collection of specific data items. Ensure there is no existing source for the data items. If there is an existing source, a new data collection may not be needed.

- If the required information is not directly available, determine whether you can derive or estimate it using existing data sources.
- If existing federal data collection programs partially meet the data requirements, determine whether the existing data programs can be altered to meet the data requirements through, for example, an inter-agency agreement.

Guideline 2.1.4: Documentation. Document the current data program objectives and make the documentation available to the public unless that documentation is restricted. Also document the updating process and include how user information is collected.

2.2 Data Requirements

Data requirements define what data are in the program and the required quality of that data. Define data requirements based on indicators designed to address the data program objectives. An indicator is a measurement which can be used to track changes in a program. The following are examples of indicators paired with an objective they address:

Objective	Indicator
Track trends in transportation services output	Transportation services index of weighted freight and passenger movement produced monthly
Provide reliable and efficient transportation for movement of goods	Hours of freight delay on the national highway system
Track progress towards increasing safety on the national highway system	Number of deaths on the national highway system

Maintaining the link from data program objectives through indicators to data requirements will help to ensure relevance of the data to users.

GUIDELINES

Guideline 2.2.1 Indicators. Each data program objective should have one or more indicators that need to be measured or capture an important metric for understanding the state of the nation. The indicators should be measurements which, when changing favorably, suggest progress toward achievement of an objective. For example, measures of GDP are an economic indicator. Some important metrics are not associated with an objective but are critical knowledge about the nation such as demographic data from the Decennial Census.

Guideline 2.2.2 Data Requirements and Indicators. Develop data requirements needed to quantify the indicators and metrics. Example: For HPMS, the indicator, "the annual vehicle miles of travel on the interstate system & other principal arteries" can lead to a data requirement for state-level measures of annual vehicle-miles traveled accurate to within 10 percent at 80 percent confidence.

Guideline 2.2.3 Describe data in detail. Each type of data should be described in the data requirements. Key variables should include requirements for accuracy, timeliness, and completeness. The accuracy should be based on how the measure will be used.

Example: For FARS, the concept, "The safety of people and pedestrians on the highways of the U.S." can lead to data requirements for counts of fatalities, injuries, and motor vehicle crashes on U.S. highways and streets. The fatalities for a fiscal year should be as accurate as possible (100% data collection), available within three months after the end of the fiscal year, and as complete as possible. The injury counts in traffic crashes for the fiscal year totals should have a standard error of less than 6 percent, be available within three months after the end of the fiscal year and have an accident coverage rate of at least 90 percent. For detailed data descriptions for dissemination, please refer to section 5.?? on Data Dictionaries.

Guideline 2.2.4 Consider standardization with other databases. To allow data comparisons across databases, use standard names, variables, numerical units, codes, and definitions. First, consider measures used for similar concepts in other DOT databases. Second, consider measures for similar concepts in databases outside DOT (e.g., The Census).

Use coding standards and make them part of the data requirements. Such standardization leads to coherence across datasets. Use codes and classifications consistent with the federal coding standards, if applicable. If a federal coding standard does not exist, consult with subject area experts to determine if applicable non-federal standards exist. Provide crosswalk tables to the federal standard codes for any legacy coding that does not meet the federal standards.

Current federal standard codes include (but are not limited to):

- Statistical Areas. OMB defines Metropolitan Statistical Areas, Micropolitan Statistical Areas, Combined Statistical Areas, and New England City and Town Areas for use in Federal statistical activities. These areas, as well as principal cities, are updated annually to reflect changes in population estimates. *(1)*

- NAICS Codes. The North American Industry Classification System (NAICS) should be used to classify establishments. NAICS was developed jointly by the United States, Canada, and Mexico to provide new comparability in statistics about business activity across North America. (NAICS coding replaced the U.S. Standard Industrial Classification (SIC) system.) *(2)*
- SOC Codes. The Standard Occupational Classification (SOC) system should be used to classify workers into occupational categories for the purpose of collecting, calculating, or disseminating data. *(3)*
- Race and Ethnicity. Classification of race and ethnicity, as well as methods of collection, should comply with OMB's Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. *(4)*
- Aviation. The International Air Transport Association, an airline industry association, establishes standard codes for airlines and airport locations *(5)*. The BTS Office of Airline Information also develops and maintains Aviation Support Tables *(6)* that provide standard codes and other information for air carriers (U.S. and foreign), worldwide airport locations, and for aircraft types and models. The BTS codes do not always agree with IATA coding.
- Standard Classification of Transported Goods (SCTG) Reporting System Codes. The SCTG coding system was created by the U.S. and Canadian governments and is used to address statistical needs regarding the transportation of products. *(7)*
- Injury Codes. "The International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)" is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. The V series codes refer to transportation-related injuries. *(8)*

Guideline 2.2.5 Standard Definitions. Coherence is one dimension of the FCSM data quality framework and includes maintaining common definitions to maintain consistency and comparability. Consult other data collections to identify common definitions to incorporate into the data collection.

Guideline 2.2.6 Documentation. Document the current data requirements and clearly post the documentation with the data.

REFERENCES

1. OMB. (2021) Metropolitan and Micropolitan. https://www.census.gov/programs-surveys/metro-micro.html
2. Census. (2022) North American Industry Classification System. https://www.census.gov/naics/
3. BLS. (2018) Standard Occupational Classification. https://www.bls.gov/soc/
4. OMB. (2016) Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity. https://www.federalregister.gov/documents/2016/09/30/2016-23672/standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity
5. IATA. (N.d) Airline and Location Code Search. https://www.iata.org/en/publications/directories/code-search/
6. BTS. (2021) Aviation Support Tables. https://rosap.ntl.bts.gov/view/dot/58890
7. Census. (N.d) SCTG Commodity Code List. https://bhs.econ.census.gov/bhsphpext/brdsearch/scs_code.html
8. CDC. (2022) ICD-10-CM Search. https://icd10cmtool.cdc.gov/?fy=FY2022

2.3 Sources of Data

There are a wide variety of sources of data. Once data requirements are set, assess the variety of options for whether they can meet the data requirements. Consider sources of data that do not require a new full-scale data collection effort before committing to a new data collection.

Using existing data is by far the most efficient approach to data acquisition. Sources of existing data can be current data programs or administrative records. "Administrative records" are data that are created by government agencies to facilitate the operations and management of an agency, but do not directly document the performance of mission functions. Besides providing a source for the data itself, administrative records may also provide information helpful in the design of the data collection process (e.g., sampling lists, stratification information). For example, state driver's license records, social security records, boat registration records, mariner license records. Another type of existing data source is a reporting collection in which the target group automatically sends data. Law or regulation dictate most of these. For example: 46 USC Chapter 61 specifies a marine casualty reporting collection, while 46 CFR 4.05 specifies details.

Another method, less costly than developing a new data collection program, is to use existing data collections modified to address your needs. The owner of such a program might add additional data collection or otherwise alter the collection process to gather data that will meet data requirements.

Another option is collection through third-party sources. Third party sources are people, businesses, or even government entities that have knowledge or collect information on the target group. Review of the data quality will be necessary before committing to using the data.

GUIDELINES

Guideline 2.3.1 Explore existing data. Research whether government and private data collections already have data that meet the data requirements. Consider surveys, reporting collections, and administrative records. If an existing collection partially meets the data requirements, work with the owner to explore adjustments to obtain the needed data.

Guideline 2.3.2 Ensure it is possible to access the entire target group. Whatever data source the department chooses, it is important that it is possible to access the entire target group. A sample approach will not include the entire target group, but all members should have a non-zero (and known) probability of selection, or the sampling will not necessarily be representative of the target group.

Guideline 2.3.3 Evaluate third-party data. Evaluate data that is acquired from external sources to assess the quality for the intended use. Obtain the highest quality version of the external data available from the source.

- Verify that the data set is the latest version, and that no revised data are available. Keep a backup copy of the data.
- Obtain the most complete data documentation available for the corresponding time periods.
- Evaluate data from external sources for data quality before deciding whether the data are appropriate for the intended BTS use. The level of BTS's evaluation effort should depend on the

thoroughness of the external source's quality control and on the importance of the data for the intended use.

Guideline 2.3.4 Confidential External Data. If the external data are confidential or proprietary, written agreements to acquire the data must stipulate the confidentiality requirements for protecting it. For further information on confidentiality of external data, please refer to the latest BTS confidentiality guide.

Guideline 2.3.5 Documentation. Document the choices made for sources and their connection to the data requirements and clearly post the documentation with the data or disseminated data products. If redesigning an existing data program, analyze and document the potential impact of changes in key variables or data collection procedures.

2.4 Data Collection Design

The design of data collection is one of the most critical phases in developing a data program. The accuracy of the data and of estimates derived from the data heavily depend on the design of data collection. A census is an ideal approach, but is not always feasible due to cost, time, respondent burden, and other resource restrictions. A probability sample is an efficient way to select a data source that is representative of the target group. When sampling people, businesses, and/or things, sampling frames (list of members of the target group) are required to select the sample. Availability of sampling frames often restricts the methods used in data collection.

GUIDELINES

Guideline 2.4.1: Target Population and Frames. Sampling frames are required to obtain information from the target population. When a data program needs a new frame, develop and implement a plan for constructing the frame.

The plan should cover:

- Choice of the target population and the rationale,
- Any exclusions applied to target and/or frame populations by design,
- Sources of lists of target population units,
- Identification and description of other frame files which exist and whether portions of other frame files will be used to construct a new file,
- Description of any multistage sampling, such as geographic area sampling, undertaken prior to development of lists of units and the stages in which the final lists will be developed,
- Methods for matching and merging population lists, if applicable,
- Data items needed for units in the frame,
- Anticipated coverage of the target population by the frame,
- Coverage rates above 95 percent overall and for each major target population subgroup are desirable.
- Consider using frame enhancements, such as frame supplementation or dual frame estimation, to increase coverage.
- If the anticipated coverage falls below 85 percent, evaluate and document the potential for bias.
- Any estimation techniques used to improve the coverage of estimates, such as post-stratification procedures,

- limitations of the frame including the timeliness of the frame, and
- Projected frequency of frame updates.

Guideline 2.4.2: Sample Design. A 100 percent data collection may be required by law, necessitated by accuracy requirements, or equally as easy/ inexpensive (e.g., data readily available, small population). Otherwise, the data collection design should include appropriate sampling methods. Any sample design chosen should ensure the sample will yield the data required to meet the objectives of the data collection. Use probability sampling so that sampling error can be estimated. Any use of nonprobability sampling methods (e.g., cut-off or model-based samples) must be explained and the sources of bias and errors must be documented. As much as possible with nonprobability sample, quantify the estimation error. The sample design should include:

- Identification of the sampling frame and the adequacy of the frame,
- The sampling unit used (at each stage if multistage design),
- Set estimation precision
- Criteria for stratifying or clustering,
- Sampling strata,
- Sample size by stratum,
- Expected responses by stratum,
- Sample selection procedures,
- The known probability (or probabilities) of selection,
- Estimated effectiveness at reaching desired precision (I.e., efficiency),
- Power analyses to determine sample sizes and effective sample sizes for key variables by reporting domains (where appropriate),
- Response rate goals,
- Estimation and weighting plan,
- Variance estimation techniques appropriate to the sample design,
- Expected precision of estimates for key variables, and
- References for the sampling methods used.

For nonprobability sample designs, include a detailed selection process and demonstrate that units not in the sample are impartially excluded on objective grounds.

Discuss potential non-sampling errors, including reporting errors, response variance, measurement bias, nonresponse, imputation error, and errors in processing the data. Indicate steps to be taken to minimize the effect of these problems on the data.

Guideline 2.4.3 Contact Strategy. Develop a contact strategy that covers initial contact and nonresponse follow-up. This strategy should address how and when the contact will be made.

Guideline 2.4.4 Data Collection Period. Establish a data collection period that allows sufficient response time for data providers to supply reliable data, including time to follow up on missing data, and meets the required dissemination schedule and mandated data collection frequency.

Guideline 2.4.5 Training. Establish a formal training process for persons involved in interviewing, observing, or reporting data to ensure that they follow the intended procedures. Train people to monitor the data collection process to ensure data quality.

Guideline 2.4.6 Expert Review. If the data collection process performed by DOT uses sampling, a statistician or other sampling expert should review the design.

Guideline 2.4.7 Plan the full data collection process prior to beginning collection

- Develop a plan for obtaining consent and confidentiality protection during sampling, data collection, processing, data analysis, and dissemination. See the most recent BTS confidentiality guide.
- Develop plans for data processing, including data editing and imputation, see Ch. 4 for guidelines
- Plan for quality assurance during each phase of the data collection process to permit monitoring and assessing the performance during implementation. Include contingencies to modify the procedures if critical requirements (e.g., for the response rate) are not met.
- Plan for evaluating data collection and processing procedures, results, and potential biases.
- Develop general specifications for an internal project management system for the complete data collection cycle that identifies critical activities and key milestones that will be monitored, and the time relationships among them.

Guideline 2.4.7 Data Collection Redesigns Changes to data collection systems can make conducting analyses over time challenging. Thoroughly assess any potential changes for their ability to address the need that motivated the changes. Conduct testing such as content and cognitive testing. Gain input from relevant stakeholders. Initially, if possible, conduct the new method in parallel with the original method. Clearly document the changes and effects of the changes.

Guideline 2.4.8 Documentation. Document the collection design and its connection to the data requirements and clearly post it with the data or data products. The documentation should include references for the sampling theory used. If the data program uses third party data collected using sampling, collect sample design information and provide it with the collection design documentation, when available

2.5 Data Program Planning Documentation

Planning activities must include the documentation of user needs and design decisions as well as the preparation of required administrative documents.

GUIDELINES

Guideline 2.5.1 Documentation of Data Needs. After establishing the data needs and requirements, prepare a detailed technical document that describes the goals and objectives of the data collection, including:

- A summary of the consultations with major data users and data providers, plus any other sources consulted,
- The information needs that will be met, including the desired accuracy, timeliness, and dissemination format(s) for the data, and
- The choices made for meeting data needs and their relationship to the requirements.

Guideline 2.5.2 Target Population and Frames Documentation. Describe the target populations and associated frames (lists of population units) in detail. Include a discussion of coverage issues (Guideline 2.4.1).

Guideline 2.5.3 Sample Design Documentation. If sampling is part of the data collection design, prepare a detailed description of the sample design (Guideline 2.4.2) and how it will yield the data required to meet the objectives of the data collection. When a non-probabilistic sampling method is employed, the survey design documentation should include:

- A discussion of what options were considered and why the final design was selected,
- An estimate of the potential bias in the estimates, and
- The methodology to be used to measure estimation error.

Guideline 2.5.4 Collection and Processing Methodology Documentation. Document the collection design and its connection to the data requirements (Section 2.4). The documentation should include the methods of obtaining data, copies of the data collection instrument and instructions, pretest design and findings, and plans for disseminating the results of the data collection to the public.

Guideline 2.5.5 Documentation of External Data Sources. All data that BTS acquires from external sources must have adequate levels of documentation. Documentation for external sources should include:

- The organization providing data,
- The exact name of the data source,
- If the data were obtained from a publication, the full publication information and source for the data within the publication,
- If the data were acquired as a data file, how the file was obtained, the date
- obtained, and the cost (if any),
- If the data were obtained from the web, the web address and the date acquired,
- The best documentation available from the external data source on the data collection design (including sampling, if used), the data collection and processing procedures, any analysis or modeling performed, and any evaluations of the data quality,
- Information on whether the external data are confidential or proprietary, and if so, a copy of the written agreement used to obtain the data,
- Any additional notes on the interpretation and use of the data,
- Any personal communications required to obtain the data, information about the data source, or information about data quality, and
- Contact information for further questions.

Guideline 2.5.6 Administrative Documents. Comply with the following requirements as part of the data collection planning and design:

- When planning and design is in its initial stages, prepare a project plan specifying schedules and resource requirements in the format specified by BTS management.
- Data collections (and related activities such as focus groups, cognitive interviews, pilot studies, field tests, etc.) are all collections of information subject to the requirements of the Paperwork

Reduction Act of 1995 (P.L. 104-13, 44 U.S.C. 3501 et seq.) and OMB's regulations (5 CFR Part 1320, Controlling Paperwork Burdens on the Public). OMB approval is required before the agency may collect information from ten or more persons outside the Federal government in a twelvemonth period. The documentation specified in this section can all be used in Part B of the submission to OMB.

- Projects that require a new IT investment or significant modification of an existing IT investment must go through the Capital Planning and Investment Control process.
- Contracts should include language stating that the contractor shall comply with all standards and guidelines contained in the BTS *Statistical Standards Manual* and the BTS *Confidentiality Procedures Manual*.

Guideline 2.5.7 Documentation of Changes to a Data Program. Any changes to a data program should be documented. Include:

- What the changes were
- Which data items are affected
- How changes affect comparison to previous data

Chapter 3 – Data Acquisition

Data collection includes all the processes involved in implementing a planned design to acquire data. Common types of data collections/Acquisitions include:

- Regulatory data collections (e.g., the airline traffic data required by 14 CFR 241),
- Administrative data collections (e.g., the border crossing data),
- Surveys (e.g., the Commodity Flow Survey), and
- Third party collections (e.g., ATRI GPS data).

This chapter contains standards for conducting data collection (Section 3.1). data acquisition from external data sources (Section 3.2), avoiding missing data (Section 3.3) and documenting the data collection/acquisition process (Section 3.4). Selection of data sources and data collection design are covered in Sections 2.3 and 2.4, respectively.

3.1 Data Collection Operations

Data collection methods can have a high impact on data quality. Collection instruments are mediums through which data are gathered. Examples include questionnaires, administrative records, web scraping, and GPS devices. The data collection method should be valid and appropriate to the data complexity, collection size, data requirements, and amount of time available. For example, a large survey requiring a high response rate, such as the Decennial Census, will often start off with a mail out, followed by telephone contact, and finally by a personal visit.

GUIDELINES

Guideline 3.1.1 Instruments and Instructions. Design the data collection instrument in a manner that maximizes data quality, while minimizing respondent burden:

- Use instrument formats that are appropriate for the method of data collection. For example, if using a self-administered, online survey collection instrument, use skip patterns whenever possible to reduce burden.
- Develop clearly written instructions to help respondents minimize missing data and measurement error.
- Require that data items are clearly defined in terms the respondents understand, with entries in a logical sequence and with reasonable visual cues and instrument formatting (if applicable). Pretest to identify problems with interpretability.
- Structure the order and presentation of survey questions such that responses do not unduly influence responses to subsequent items.
- Minimize the number of data calculations and conversions the respondent must make.
- For computer-assisted and other forms of electronic data collection (using GPS devises, sensors, etc.):
    - Test for validity and reliability under similar conditions to those of the planned data collection.

  - Develop protocols for the backup and recovery of data.
  - If possible, have alternate methods of data collection available in case of equipment failure. Otherwise, develop plans to impute or adjust for faulty or missing observations.
- Establish protocols that minimize measurement error, such as re-interviewing to capture response measurement error, establishing follow-up periods that are reasonable for personal data collections, and developing computer systems that ensure internet data collections function properly.

Guideline 3.1.2 Pretesting. For new data collections or major revisions of ongoing collections, pretest all components so that they minimize measurement error and function properly prior to full implementation.

- One component of pretesting is testing components of a data collection prior to a field test (for example, cognitive testing and/or calibration studies).
- Another component of pretesting is a pilot test. Before full-scale data collection, pilot test components of a data collection that previous work cannot successfully demonstrate. The design of a pilot test should reflect realistic conditions, including those likely to pose difficulties for the data collection.

Guideline 3.1.3 Status Tracking. Use a status tracking procedure to track the progress of data collection. Examples of trackable data for a survey-based data collection include when and who have been contacted, logins, how long respondents are in the system, how long respondents spend on each page, and response rates. These data can help inform the burden and effectiveness of the data collection.

Guideline 3.1.4 Information Collection Request. Provide respondents with an Information Collection Request (ICR) when collecting information. The ICR summary is usually placed on the information collection instrument. When data are collected under a pledge of confidentiality, explain the pledge and how it will be enforced to the respondent. Refer to agency confidentiality guidelines such as the latest version of the BTS *Confidentiality Procedures Manual*. For further reading on ICRs, please review the paperwork reduction act guide at https://pra.digitial.gov/.

Guideline 3.1.5 Protecting Confidential Data. In all phases of data collection, protect confidential data from unauthorized access or release. For more detail, consult agency confidentiality guidelines such as the *BTS Confidentiality Procedures Manual.*

- Protect identifying information of respondents as collected or on the sample frame from unauthorized release or access.
- Ensure that controls are in place to prevent unauthorized access to electronic information collections.
- Ensure that all data collection staff have received confidentiality training and signed a non-disclosure form prior to collecting data.
- Use secure means when handling and storing the data during collection to protect against disclosure.
- Use other means to protect confidential information as outlined in the latest version of the *Confidentiality Procedures Manual*.

Guideline 3.1.6 Documentation. Document the collection operation procedures and post the documentation with the data and data products. If using third party data collection, provide documentation on procedures used by the third party as well.

3.2 Data Acquisition Process and Checking

When using third party sources, check credibility of the data producer and vet their data products as thoroughly as possible. Check if they have introduced some degree of error in their collection processes. When a third-party data collector supplies data, some initial data check and follow-up for missing data will dramatically reduce the incidents of missing data.

GUIDELINES

Guideline 3.2.1 Data Checking When acquiring data from an external source, perform big picture reasonableness checks. Examples of checks to use if applicable include:

- Format check – Confirm the data are in a usable and consistent format
- Value check – Confirm data values are within the expected range of values and identify any missing or duplicate data.
- Trend check – Confirm that the data follow a reasonable trend and continue any trends from earlier data
- Comparison check – Compare the data to similar data if available

Guideline 3.2.2 Documentation Document all data checking methods used, and track any missing values or changes made to the data.

3.3 Missing Data Avoidance

Some missing data occur in most data collection efforts. Unit-level missing data occur when a report is completely missing or is received and cannot be used (e.g., garbled data, missing key variables). Item-level missing data occur when data are missing for one or more items in an otherwise complete report. For example, for an incident report for a hazardous material spill, unit-level missing data occur if the report was never sent in, or all entries were unusable. Item-level missing data would occur if the report was complete, except it did not indicate the quantity spilled.

The extent of unit-level missing data can be difficult to determine. If a report should be sent in whenever a certain kind of incident occurs, then non-reporters can only be identified if crosschecked with other data sources. On the other hand, if companies are required to send in periodic reports, the previous period may provide a list of the expected reporters for the current period. This can also be true for item-level missing data. For example, in a travel survey asking for trips made, forgotten trips may not be known.

Some form of missing data follow-up will dramatically reduce the incidents of both unit-level and item-level missing data. A process to recontact the data source can be used, especially when they left out critical data. Incident reporting collections can use some form of cross-check with other data sources to detect when incidents occur but are not reported. If using electronic data collection, alert respondents to any missing information before submitting.

GUIDELINES

Guideline 3.3.1 Quality Assurance. Develop protocols to monitor data collection activities, with strategies to identify and correct problems to ensure quality during data collection:

- Implement a process control system during data collection to monitor data quality. Integrate the quality control system into the data collection process and ensure it enables staff to identify and resolve problems. The control system should also provide data quality measurements for use as indicators of data collection performance and data quality.
- All data collection programs should have some follow-up of missing reports and data items, even if third-party sources provide the data. For incident reporting collections where missing reports are challenging to track, some form of checking process should exist to reduce missing reports.
- Use a verification process in data entry to ensure entry errors remain below a set limit based on data accuracy requirements. Include data verification rules in online or other electronic data collection systems.
- Conduct refresher training periodically for persons involved in interviewing, observing, or providing data to maintain proper procedures and standards set out in the data collection design.
- Track on-going response rates and item nonresponse for key variables. Conduct an evaluation of potential item nonresponse bias if response rates fall below 80 percent for core items.
- Determine the core items to obtain when a respondent is unwilling to complete the whole information collection instrument. Target the core items to meet the minimum standard for unit response and to analyze nonresponse bias.

Guideline 3.3.2 Encourage Cooperation. To encourage data providers and respondents to participate, train data collection staff on obtaining cooperation, building rapport, and converting refusals, even for mandatory data collections. Means such as prenotification letters, multiple contacts, and reminder notices can improve response rates and data quality.

Guideline 3.3.3 Documentation. Document the missing data avoidance procedures and clearly post them with the data and data products. Documentation should address how the collection process was designed to produce high rates of response. If a third party collects the data, the data collection program documentation should indicate how the third-party deals with missing data if that documentation is available.

3.4 Documentation of Data Collection Procedures

The data collection procedures should be documented both for internal staff reference and for the public. Documentation should be thorough enough to allow reproduction of the steps leading to the results.

Guideline 3.4.1: Documentation of Data Collection Operations. The data collection operations documentation should include:

- The method of data collection (e.g., mail, telephone, Internet, etc.), including methods used to track and follow up delinquent reports,
- The data collection period, response rate achieved by the end of the period, and final response rate achieved,
- Copies of materials used in the data collection, including instructions given to data providers,
- Copies of materials used in training data collection and data provider staff,
- Schedule of data collection operations,
- Any response analysis or other validation surveys conducted for new data collection efforts,
- Quantification of response errors to the extent possible
- Non-response bias analysis when warranted.

Chapter 4 - Processing Data

Once data are collected, process it to mitigate errors and prepare data products.

4.1 Data Protection

Take precautions throughout data processing to protect the data from disclosure, theft, or loss.

GUIDELINES

Guideline 4.1.1: Confidentiality Procedures. Implement the confidentiality procedures in the latest agency confidentiality guidelines such as the BTS *Confidentiality Procedures Manual* to protect the data from unauthorized disclosure or release during data production, use, storage, transmittal, and disposition.

Guideline 4.1.2: Security of Information Systems. Follow the information system security procedures in the latest agency confidentiality guidelines such as the BTS *Confidentiality Procedures Manual*, and periodically monitor and update them. Ensure that:

- Data files, networks, servers, and computers are secure from malicious software, unauthorized access, or theft.
- Access to confidential data is controlled so that only authorized staff can read and/or write to the data. The project manager responsible for the data should periodically review staff access rights to guard against unauthorized release or alteration.

Guideline 4.1.3: Data Storage. Develop and implement routine data backups. Secure backup data from unauthorized access or release.

4.2 Data Editing and Coding

Data editing is the application of checks that identify missing, invalid, duplicate, inconsistent entries, or otherwise point to data records that are potentially in error. Typical data editing includes range checks, validity checks, consistency checks (comparing answers to related questions), and checks for duplicate records. Editing is a final inspection-correction method. It is almost always necessary, but data quality is better achieved much earlier in the process through clarity of definitions, forms design, data collection procedures, etc.

Coding means to add codes to the data set as additional information or converting existing information into a more useful form. Codes may indicate information about the collection, conversions of data, such as text data, into a form more useful for data analysis, or editing and missing data actions taken.

GUIDELINES

Guideline 4.2.1: Types of Edits. Apply an editing process to every data collection and to third-party data to reduce error in the data. The editing process must include checking and correcting for the items below:

- Omission or duplication of records/units,
- Data that fall outside a pre-specified range, or for categorical data, data that are outside the specified categories,
- Data that contradict other data within an individual record/unit,
- Data inconsistent with past data or with data from outside sources,
- Missing data that can be filled from other portions of the same record or through follow-up with the data provider, and
- Incorrect flow through prescribed skip patterns.

Guideline 4.2.2: Editing Process. In a data editing system:

- Develop editing rules before any data processing. Rules may be modified during data processing.
- Minimize manual intervention because it can cause inconsistent applications of the edit rules and human error.
- Set the acceptable data ranges for outlier checks at broad enough levels so that legitimate special effects, trend shifts, or industry changes are not erroneously removed.

Guideline 4.2.3: Edit Resolution. Several actions are possible when a data value fails an edit check. Recommended procedures are:

- Verify with the original source or respondent and correct,
- Change the data value to the most likely value based upon other information collected,
- Impute a substitute value,
- Replace the failed value with a missing value indicator (Guideline 4.4.2), or
- Accept the data value as reported. Provide reasons for overriding edits.

For administrative or regulatory data, any changed value needs the data provider's acceptance. Notify the source if a change is made to data provided by an external source.

Guideline 4.2.4: Codes for Missing and Inapplicable Data. Use codes on the file that clearly distinguish between cases where an item is missing and cases where an item does not apply, such as when skipped by a skip pattern.

- Distinguish between data missing initially from the source, unreadable data, and data deleted in the editing process.
- If the data collection instrument contains skip patterns, distinguish between items skipped and items not ascertained (such as refusals).
- Do not use blanks and zeros to identify missing data, as they are confused with actual data. Similarly, do not use numeric codes like a series of nines or eights for missing numeric items if these could be legitimate reported values.
- If the source did not code a data file, the level of coding effort should depend on how BTS plans to use the file and on whether BTS plans to further disseminate the file.
- For data in tabular form, the *BTS Guide to Style and Publishing Procedures* contains several symbols and abbreviations to place in cells with various types of missing or inapplicable data.

Guideline 4.2.5: Indicating Edit Actions and Imputations. Code the data set to indicate edit actions and imputed values.

- Indicate whether cases passed or failed each edit. If a case fails an edit, indicate the edit disposition.
- If more than one method could impute a missing data item, indicate the imputation method used.

Guideline 4.2.6: Coding Text Information. Although it is preferable to pre-code responses, it may be necessary to code open-ended text fields for further use.

- To code text data for easier analysis, use standardized codes if they exist. Develop other types of codes by using existing DOT or other federal agency practice, or by using standard codes from industry or international organizations, when they exist.
- When manually coding text, create a quality assurance process that verifies at least a sample of the coding to determine if a specific level of coding accuracy and reliability is being maintained.

4.3 Handling Missing Data

Untreated, missing data can introduce substantial error into estimates. Frequently, there is a correlation between the characteristics of those missing and those reported, resulting in biased estimates. For this reason, it is often best to employ adjustments and imputation to mitigate this damage.

One method used to deal with unit-level missing data is weighting adjustments. All cases, including the missing cases, are put into classes using variables known for both types. Within the classes, the weights for the missing cases are evenly distributed among the non-missing cases. A second method is to use imputation. Imputation is a process that substitutes values for missing or inconsistent reported data. Such substitutions may be strongly implied by known information or derived as statistical estimates. If imputation is employed and flagged, users can either use the imputed values or deal with the missing data themselves.

The impact of missing data for an estimate is a combination of how much is missing and how much the missing data differ from the available data in relation to the estimate (usually unknown). For example, given a survey of airline pilots that asks about near-misses they are involved in and whether they reported them, it is known how many of the sampled pilots did not respond. You will not know if the ones who did respond had a lower number of near-misses than the ones who did not.

GUIDELINES

Guideline 4.3.1: Basis for Rates. Calculate unit and item response rates based either on the probability of selection (for household or personal data collections) or on the unit's measure of size for industry or establishment data collections.

- Base proportions of the total industry on a measure of size available for all eligible units (e.g., annual operating revenue, total employment).
- For sample surveys, use the inverse of the probability of selection (base weights) in response rate calculation. For 100 percent (universe) data collections, the base weight for each unit is one.
- For sample designs using unequal probabilities, such as stratified designs with optimal allocation, report weighted missing data rates along with unweighted missing data rates.
- If sample substitutions were made, calculate response rates without the substituted cases.

Guideline 4.3.2: Unit Response Rates. Calculate unit response rates (*RRU*) as the ratio of the number of completed data collection cases (*CC*) to the number of in-scope sample cases. Multiple different categories of cases comprise the total number of in-scope cases:

CC = number of completed cases;

R = number of cases that refused to provide any data;

O = number of eligible units not responding for reasons other than refusal;

NC = number of noncontacted units known to be eligible;

U = number of units of unknown eligibility; and

e = estimated proportion of units of unknown eligibility that are eligible.

The unit response rate represents a composite of these components:

$$RRU = \frac{CC}{CC + R + O + NC + e(U)}$$

- Complete cases may contain some missing data items. Data collection staff and principal data users should jointly determine the criteria for considering a case to be complete.
- The denominator includes all original survey units that were identified as being eligible, including units with pending responses with no data received, new eligible units added to the data collection effort, and an estimate of the number of eligible units among the units of unknown eligibility. The denominator does not include units deemed out-of-business, out-of-scope, or duplicates.
- An unweighted version of the unit response rate can be used for tracking and analyzing data collection operations.
- A simple way to calculate *e(U)* is to compute the weighted ratio of eligible to ineligible in completed cases or eligibility-known cases and assume the same ratio will apply to the *U* cases.
- If a data collection has special circumstances that justify a formula other than the one above, such as longitudinal or partial response considerations, use a more appropriate formula if accompanied by a full explanation of the calculation method.
- When a data collection has multiple stages, calculate the overall unit response rates (*RROC*) as the product of two or more unit-level response rates.

Guideline 4.3.3: Item Response Rates. Calculate item response rates (*RRI*) as the ratio of the number of respondents for whom an in-scope response was obtained (CC_x for item x) to the number of respondents who were requested to provide information for that item. The number requested to provide information for an item is the number of unit level respondents (*CC*) minus the number of respondents with a valid skip for item x (V_x). When an abbreviated questionnaire is used to convert refusals, the eliminated questions are treated as item nonresponse.

$$RRI_x = \frac{CC_x}{CC - V_x}$$

- Calculate the total item response rates (RRT_x) for specific items as the product of the overall unit response rate (*RRO*) and the item response rate for item x (RRI_x).

$$RRT_x = RRO * RRI_x$$

Guideline 4.3.4: Multiple Samples. When calculating a response rate with supplemented samples, base the reported response rates on the original and the added sample cases. However, when calculating response rates where the sample was supplemented during the initial sample selection (e.g., using matched pairs), calculate unit response rates without the substituted cases included (i.e., only the original cases are used).

Guideline 4.3.5: Unit Nonresponse Bias. Given a survey with an overall unit response rate of less than 80 percent, conduct an analysis of nonresponse bias using unit response rates as defined above, with an assessment of whether the data are missing completely at random. As noted above, the degree of nonresponse bias is a function of not only the response rate but also how much the respondents and nonrespondents differ on the survey variables of interest. For a sample mean, an estimate of the bias of the sample respondent mean is given by:

$$B(\bar{y}r) = \bar{y_r} - \bar{y_t} = \left(\frac{n_{nr}}{n}\right)(\bar{y_r} - \overline{y_{nr}})$$

Where:

$\bar{y_t}$ = the mean based on all sample cases;

$\bar{y_r}$ = the mean based only on respondent cases;

$\overline{y_{nr}}$ = the mean based only on the nonrespondent cases;

n = the number of cases in the sample; and

n_{nr} = the number of nonrespondent cases.

For a multistage (or wave) survey, focus the nonresponse bias analysis on each stage, with particular attention to the "problem" stages. A variety of methods can be used to examine nonresponse bias, for example, make comparisons between respondents and nonrespondents across subgroups using available sample frame variables. In the analysis of unit nonresponse, consider a multivariate modeling of response using respondent and nonrespondent frame variables to determine if nonresponse bias exists. Comparison of the respondents to known characteristics of the population from an external source can provide an indication of possible bias, especially if the characteristics in question are related to the survey's key variables.

Guideline 4.3.6: Item Nonresponse Bias. If the item response rate is less than 70 percent, conduct an item nonresponse analysis to determine if the data are missing at random at the item level for at least the items in question, in a manner like that discussed in Guideline 4.3.5.

Guideline 4.3.7: Not Missing at Random. In those cases where the analysis indicates that the data are not missing at random, the amount of potential bias should inform he decision to publish individual items.

Guideline 4.3.8: Imputation. Decisions regarding whether to adjust data, adjust weights, and impute for missing data should be based on how the data will be used and the assessment of bias due to missing data.

- To avoid biased estimates, include imputed data in any reported totals.
- When used, imputation procedures should be internally consistent, based on theoretical and empirical considerations, appropriate for the analysis, and based on the most relevant data available.
- Since most data sets are subject to analysis by users to detect relationships between variables, implement imputation methods that preserve multivariate relationships.
- To ensure data integrity, re-edit data after imputation.
    - If imputation is used, add a separate field containing a code (i.e., a flag) to the imputed data file indicating which variables have been imputed and by what method.

Guideline 4.3.9: Weight Adjustments. For data collections involving sampling, adjust weights for unit nonresponse, unless it warrants unit imputation. Adjust weights for missing units within classes of sub-populations to reduce bias. For sample designs using unequal probabilities (e.g., stratified designs with optimal allocation), report weighted missing data rates along with unweighted missing data rates.

4.4 Production of Estimates and Projections

An estimate approximates some characteristic of the target group, such as the average age. A projection is a prediction of a measure of the target group, usually in the future. Example: The average daily traffic volume at a given point of the Garden State Parkway in New Jersey two years from now.

GUIDELINES

Guideline 4.4.1 Derived Data: Use derived data to enhance the data set without additional burden on data suppliers. For example, the data collection can note the departure and arrival airports, and the distance of the flight can be calculated after collection.

Guideline 4.4.2 Weights: Weights should be used in all estimates from samples. Weights give the number of cases in the target group that each case represents and are calculated as the inverse of the sampling probability. If using weights, adjust weights for nonresponse as discussed in section 4.2.

For example, the National Household Travel Survey is designed to be a sample representing the households of the United States, so the total of the weights for all sample households should equal the number of households in the United States. Due to sampling variability, it won't. Since we have a very good count of households in the United States from the Decennial Census, we can do a ratio adjustment of all weights to make them total to that count.

Guideline 4.4.4 Estimates and Projections: Construct estimation methods using published techniques or your own documented derivations appropriate for the characteristic being estimated. Consult forecasting experts when determining projections.

Guideline 4.4.5 Standard Error: Standard error estimates should accompany any estimates from samples. Calculate standard errors taking the sample design into account. For more complex sample designs, use replicated methods (e.g., jackknife, successive differences) incorporating the sample weights. Consult with a variance estimation expert. Ensure that any statistical software used in constructing estimates and their standard errors use methods that can account for the design of the data collection.

4.5 Monitoring and Evaluation

Monitor and evaluate each data processing activity, both to assess the impact on data quality and to inform data users.

GUIDELINES

Guideline 4.5.1: Quality Control. Establish quality control procedures to monitor and report on the operation of data processing procedures.

- Incorporate quality control into the processing procedures to automatically produce outputs useable by data system managers. Outputs produced during data processing should be used to adjust procedures for higher quality results and greater efficiency.
- Monitor failure rates for each edit and by case. Analyze the pattern of edit failures graphically to pinpoint problems more easily and prioritize items for follow-up.
- When applicable, automate referring data problems to data providers for quicker resolution.
- Maintain information on the amount of missing data, actions taken, and problems encountered during imputation for the documentation.

Guideline 4.5.2: Unit Response Analysis Requirement. Conduct an analysis of nonresponse for any data collection with an overall unit response rate less than 80 percent. The objective is to measure the impact of the nonresponse and to determine whether the data are missing at random.

- Compare respondents and nonrespondents across subgroups using external or frame data, if available, or through a nonresponse follow-up survey.
- Compare respondents' characteristics to known characteristics of the population from an external source. This comparison can indicate possible bias, especially if the characteristics in question are related to the data collection effort's key variables.
- Consider multivariate modeling of response using respondent and nonrespondent external data to determine if nonresponse bias exists.
- For a multi-stage data collection effort, focus the response analysis on the stages with the higher missing data rates.
- Evaluate the impact of weighting adjustments on nonresponse bias.

Guideline 4.5.3: Item Response Analysis Requirement. If the item response rate is less than 70 percent, conduct an item nonresponse analysis to determine if the data are missing at random at the item level, in a similar fashion to Guideline 4.5.2.

- Analyze missing data rates at the item level and compare the characteristics of the reporters and the non-reporters.

- For some data collections, such as incident data collections, missing data rates may not be known. In such cases, provide estimates or qualitative information on what is known.

Guideline 4.5.4: Timing of Nonresponse Bias Analyses. Conduct unit and item nonresponse bias analyses prior to the release of any information products

- Analyze the missing data effect at least annually if the data collection occurs more than once a year or is continuous.
- Analyze the missing data effect each time data are collected if the collection occurs annually or less often.
- For data collections from longitudinal panels, analyze the effect of missing data after each collection due to attrition of respondents.

Guideline 4.5.5: Publishable Items. In those cases where the analysis indicates that the data are not missing at random, the decision to publish individual items should be based on the amount of potential bias due to missing data.

- If the missing data bias analysis shows that the data are not missing at random and the total item missing data rate is less than 70 percent, the estimate should be considered unreliable.
- Suppress or flag estimates that are unreliable due to missing data.

4.6 Data Analysis and Interpretation

Design analyses to focus on answering the key questions. For analysis of data collected using complex sample designs, take the design into account when determining analysis methods (e.g., use weights, replication for variances). The "robustness" of analytical methods is their sensitivity to assumption violation. Robustness is a critical factor in planning and interpreting an analysis.

GUIDELINES

Guideline 4.6.1: Criteria for the Conduct of Data Analysis. The data analysis should be relevant, objective, comprehensive, and add value to existing information. To meet these goals, data analysts need to:

- Conduct the data analysis in an objective and policy-neutral manner that focuses on the statistical and economic facts.
- Maintain awareness of subject matter issues so that the data analysis can address topics of interest and importance.
- Consult with subject area specialists about relevant issues, the strengths and weaknesses of data sources, and important references to key topic elements.
- If the data analysis is not comprehensive, indicate what further types of data analysis should be considered and whether the agency plans to do that work.

Guideline 4.6.2: Data Analysis Plan Requirement. Prepare a data analysis plan prior to the start of the data analysis, ideally prior to data collection.

- Include the purpose of the data analysis, the research question, target audience, data sources (including a description and any limitations), key variables to be used, and the data analysis

methods. Also provide target completion dates and an estimate of the resources needed to complete the product.
- Subject matter experts should review the plan to ensure that the proposed data analysis will answer relevant questions. Data analysis experts should review the plan to ensure that appropriate data and methods will be used.
- The designated manager must approve the data analysis plan.

Guideline 4.6.3: Data Analysis Methods. Analyses must use theory and methods justifiable by reference to statistical literature or by mathematical derivation.

- Use appropriate analysis methods for complex sample, time series, and geospatial data, or variance estimates may be substantially biased.
- If extensive seasonality, irregularities, known special causes, or variation in trends are present in the data, take those into account in the trend analysis.
- Use robust methods if in doubt about the quality of the data (i.e., the quality of the data cleaning) or about the suitability of the data for analysis by standard parametric methods.

Guideline 4.6.4: Indicating Uncertainty. Statistical statements should be accompanied by some assessment of the limitations and uncertainty of the results.

- Estimated errors due to statistical sampling or modeling indicate the reliability of the estimate. However, these estimated errors do not account for bias, which may have a greater effect on accuracy, and does not decrease as the number of cases increases.
- Analysts must consider data quality issues related to measurement error and missing data. The purpose, design, methods, and quality of processing can all place limitations on the analysis and interpretation of the data. If possible, quantify and eliminate biasing effects. Otherwise, discuss the nature and estimated magnitude of these limitations in the report.

Guideline 4.6.5: Inference and Comparisons. Support statistical statements with proper testing and inference procedures.

- Sampling error estimates should accompany any estimates from samples.
    - For complex sample designs, the BTS office originating the data should provide guidance on estimation and variance calculation. The guidelines should cover proper use of weights and recommend a maximum coefficient of variation and a minimum cell size for usability.
- When doing multiple comparisons with the same data between subgroups, include a note with the test results indicating whether the significance criterion (Type I error) was adjusted and, if adjusted, the method used.
- Not every statistically significant difference is important. Given a comparison with a statistically significant difference, subject matter expertise is needed to determine whether the difference is important. In the measure's context and fluctuation over time, it may be considered insignificant.

Guideline 4.6.6: Bridge Estimates. If the scope of data collection changes or part of a historical series is revised, publish data for both the old and the new series for a suitable overlap period.

Guideline 4.6.7: Assumptions and Diagnostics. State all statistical assumptions (such as assumptions about data distributions or structured dependence) made during the data analysis.

- Perform diagnostics to detect violations of assumptions and provide the results of the diagnostics in the report. Plots of data and statistical output, such as residuals, are often useful in detecting violations of assumptions.
- For each assumption, include a discussion of the likelihood that the assumption will be violated by small or large amounts and the robustness of the data analysis method to each such violation.

4.7 Documentation of Data Processing Procedures

The data processing procedures must be documented for both BTS and public use. For external source data, the documentation must include procedures used by the external source as well as procedures that were implemented on the data at BTS. Documentation must allow reproduction of the steps leading to the results.

GUIDELINES

Guideline 4.7.1: Edit Procedures. Documentation must describe:

- The edit rules and their purpose,
- Procedures for handling records that fail edits,
- A description of the codes used to indicate edit disposition (Guideline 4.2.3), and
- The procedures for, and the results of, any edit performance evaluations.

Guideline 4.7.2: Measures of Edit Performance. For key edits as identified by the data collection staff, maintain measures for the number of:

- Edit messages, by edit disposition (Guideline 4.2.3),
- Edit messages resulting in revisions of the original data, and
- Edit messages overridden, by reason for overriding the edit.

Guideline 4.7.3: Procedures for Handling Missing Data. Documentation of procedures for handling missing data must include:

- The unit response rate or rates,
- Item response rates for key variables as identified by the data collection staff,
- Item response rates for any items with response rates less than 70 percent,
- Formulas used to calculate unit and item response rates,
- Results of response bias analyses,
- Full documentation of the methods of imputation or weight adjustments,
- A description of the coding schemes used to identify missing and imputed values, and
- An assessment of the nature, extent, and effects of imputation or weight adjustments.

Guideline 4.7.4: Procedures for Coding Text Information. Document both the source for any coding scheme used and the coding process (whether automated or manual) and make it available to data users. Any reliability or accuracy studies of the coding process should also be documented and made available.

Guideline 4.7.5: Derived Data Items. Documentation should include all formulas, detailed descriptions on how the item was created, and the sources of any external information used to derive additional data items for the file.

Guideline 4.7.6: Information Systems Documentation. Systems for the processing of data should have documentation of all operations (both automated and manual) necessary to operate, maintain, and update the systems.

- The documentation should provide an overview of integrated manual and automated operations, workflow, interfaces, and personnel requirements.
- Documentation should be sufficiently detailed and complete that personnel unfamiliar with the systems can become knowledgeable and operate them, if necessary.
- Information systems documentation may be incorporated into existing documentation or written as a separate document.

Guideline 4.7.7: Data Analysis Documentation. The data analysis report must contain details of the methods used during the data analysis, including a description of software used, a discussion of the data analysis assumptions, and key information relevant to obtaining the data analysis results.

- Document all methods, assumptions, diagnostics, and robustness checks. Provide references to support the methods used in the data analysis, or a derivation of the theory supporting the method used in the report.
- Include a statement of the limitations of the data analysis, including coverage and response limitations and statistical variation.
- Archive the data and models used in the data analysis so the estimates can be reproduced.
- Archive supporting technical documentation, such as standard error and significance test calculations, that help ensure transparency and reproducibility.
- For recurring reports, consider producing a methodological report.

Guideline 4.7.8: Documentation Updates. Update documentation whenever a major change to the processing system is made, but at least annually when the frequency is less than annual.

Chapter 5 - Dissemination of Information

Dissemination is the distribution of information to the public, in any medium or form, including press releases, reports, data files, or web products. OMB guidance emphasizes the need for documentation at every stage of the data lifecycle (OMB 2006). Documentation supports credibility among users and allows improvements in the data process. These standards cover releasing information (Section 5.1) and ensuring the accuracy and interpretability of different types of information products: text (Section 5.2), tables, graphs, and maps (Section 5.3), and micro data (Section 5.4). The standards also cover issues affecting all information products: rounding (Section 5.5), data protection (Section 5.6), and revisions (Section 5.7).

5.1 Releasing Information

Procedures for the release of information products to the public must receive pre-dissemination reviews (disclosure, content matter, statistical and methodological) and must include provisions for ensuring fair access to all users.

GUIDELINES

Guideline 5.1.1: Release Schedules. To provide fair access to the public, major information products should follow publicly published release schedules. Punctuality is important for meeting user expectations and reducing the appearance of political interference in a scheduled release.

- Provide the schedule for the release of information products.
- Protect information to be published against any unauthorized pre-release or disclosure in advance of the publication schedule.

Guideline 5.1.2: Ease of Accessibility and Understanding. Information products should be accessible to the public. All information products should be compliant with Section 508 of the Rehabilitation Act.

- All information products disseminated through the Internet should comply with the requirements for Section 508 of the Rehabilitation Act of 1973, as amended.
- Use codes, abbreviations, and acronyms sparingly and define them in accordance with the latest agency guidelines such as the *BTS Style Guide*. Provide definitions to the user in the product.
- As appropriate, information products should also include definitions of any subject-matter-specific or otherwise technical terms.

Guideline 5.1.3: Formal Pre-Dissemination Review. All information products require pre-dissemination review to ensure compliance with OMB and DOT Information Quality Guidelines, and agency standard procedures.

- Before sending an information product outside the originating office for review, the product manager should:
    - o Verify compliance with all applicable standards and guidelines
    - o Double-check facts,

o Proofread text, and

o Clearly mark the product as a draft for review only, and not for attribution or further distribution.

- All information products require a confidentiality protection review.
- Verify calculations through an independent recalculation of a random selection of statistics in the information product.
- Persons not directly involved in preparing the information product should proofread the text and verify that numbers in tables, graphs, maps, and text are consistent.
- All information products require a subject-matter review by someone who is familiar with the topic area and with the techniques used. The information product may require a separate review of the statistical methodology.
- Publication specialists should edit text products to ensure consistency and readability.
- Each agency has its own product review process. For BTS products, the appropriate office director should review and clear all information products before submitting the products to the Director or the Director's designee for final review the product. BTS statistical products are exempt from further review.
- Information products to be posted on the web require review for compliance with web guidelines.

Guideline 5.1.4: External Peer Review. If using an external peer review process:

- Select peer reviewers primarily based on necessary technical expertise,
- Any non-government peer reviewers paid by BTS must disclose to DOT any prior technical/policy positions they may have taken on the issues at hand and their sources of personal and institutional funding (private or public),
- Conduct peer reviews in an open and rigorous manner, and
- Consider all relevant technical comments, although outside reviews are not binding.

Guideline 5.1.5: Contact Information. All information products must include a contact reference to someone who can answer questions about the data product.

5.2 Text Discussion

Present information to the public clearly and objectively, including a full disclosure of source(s).

GUIDELINES

Guideline 5.2.1: Presentation. For style and presentation guidelines, please refer to the BTS Style Guide.

Guideline 5.2.2: Sources. Data presented in the text that do not refer directly to the tables, graphs, or maps in the text must have a source reference.

- Information used in statistical products should come from known reliable sources.
- Sources for which methodological information is unavailable (such as proprietary data) must include advisories indicating the lack of source and accuracy information.

Guideline 5.2.3: Data Discussions. Discussions of data should be objective and make statistically appropriate statements.

- Fully discuss fundamental changes within time series data collections. These changes may include, but are not limited to, changes to how the data were collected, changes in definitions, changes to the population, or changes in processing methods.
- Statistical descriptions should indicate the amount of uncertainty. Only discuss differences or changes if the appropriate statistical tests verify their statistical significance. Terms such as "confidence," "reliability," "significant," and "variance" should only be used in the statistical sense.
- Avoid statements that imply a specific cause and effect relationship where one has not been established. Speculative statements about possible causes are acceptable if worded as speculation and not fact, and if supported by legitimate research citations.
- No policy recommendations may be made regarding solutions to problems except regarding data requirements.

5.3 Tables, Graphs, and Maps

Tables, graphs, and maps in statistical products must accurately and effectively convey the information intended. As far as possible, they should be interpretable as stand-alone products in case they become separated from their original context. Methods used to produce data displayed in tables, graphs, and summary data should be available to the reader.

GUIDELINES

Guideline 5.3.1: Identifying Content and Sources. As far as possible, tables, graphs, and maps should be interpretable as stand-alone products.

- Clearly word titles for tables, graphs, and maps and answer three questions: what (data presented), where (geographic area represented), and when (date covered by data).
- All tables, graphs, and maps must have a complete source note. Include information not immediately evident from the main body of the presentation, such as definition of codes, acronyms and special terms, and anything else that would not be obvious to the general reader.
    - Detail source references sufficiently for a reader to identify the data used. Source notes in all products must give a full citation for the actual source from which the data were taken, even if that source merely collected data from other sources.
    - Note the "as of" date for the source because they may have been updated. Web links should include the URL and date accessed. Even a report featuring results entirely from one source should have the source note with each table, graph, or map, in case they are separated from the report.
- When presenting estimates that are calculated using data from external sources, note each source and add a statement describing how the calculation was done. If the calculation is complex (e.g., a weighted average constructed from raw data and weights, describe the methods used or a reference to where they are described.
- Use footnotes to clarify data illustrations, tables, graphs, and maps regarding abbreviation symbols and general notes.

Guideline 5.3.2: Consistency of Presentation. To facilitate comparability, be consistent in constructing tables, graphs, and maps within an information product.

- Tables, graphs, and maps within the same information product should use similar fonts, units, spacing, and line thicknesses. Symbols and codes should also be similar throughout an information product.
- For comparability across BTS products, tables and graphs must comply with the latest agency publication guidelines such as the *BTS Style and Formatting Guidelines*.

Guideline 5.3.3: Tables. Each cell in a table must have a number, a zero indicator, or a symbol indicating the reason that data are not displayed. Numbers in tables must comply with the following criteria:

- Use no greater precision than is warranted by the data.
- Only display zeros for values that are true zeros. If a value rounds to zero, use alternate symbols, such as "--," to indicate that the estimate rounds to zero in the units being presented.
- For sample-based zero estimates, use alternate symbols to indicate that the estimates are negligible, but possibly non-zero, in the population.
- All tables that should logically sum to either 100 percent or some other numeric total must provide a note if independent rounding or missing data affected the summation.
- Comply with the table formatting described in the BTS Style Guidelines

Guideline 5.3.4: Graphs. Design graphs to maximize clarity and comparability within the information product and with other statistical products.

- Design color graphs to show sufficient contrast if printed in black and white or viewed by a colorblind user. Web graphs need alternative text for use by screen readers.
- Graph titles and axis labels should be clear with no unexplained or undefined acronyms or industry jargon. In graphs with axes, indicate well-defined variable names and units for each axis. Label both axes of a graph with the names of variables, except where the axis label "years" is unnecessary because the years are shown.
- Graphs that users are likely to compare should have similar scaling to facilitate the comparison.
- Gridlines can be helpful to users if kept inconspicuous.
- Minimize non-data clutter.
- Minimize use of stacked bar or line graphs. They tend to present minimal information and are usually harder to interpret than simple tables or line graphs.
- Do not use 3D graphs to present two-dimensional data
- When using time intervals, spacing should be equidistant only if the intervals are equidistant.
- In graphs, a vertical numerical axis should normally include zero or a break indicator (two slashes). If adding such a break is not reasonable due to software restrictions, add a note that the vertical axis is not zero-based.

- For graphs showing relative quantities such as an index, zero is not a meaningful reference point. In such graphs, use the natural basis (such as 100) as a reference line in the graph.

Guideline 5.3.5: Statistical Maps. Statistical maps must comply with good cartographical practice. BTS personnel may refer to the BTS Cartography Style Guide for detailed guidelines.

5.4 Micro Data Releases

Where confidentiality protections permit their release, release micro data (unit-level data) in a manner that facilitates its usefulness to the public. Documentation must be readily accessible to users, provide the metadata necessary for users to access and manipulate the data, and clearly describe how the information is constructed.

GUIDELINES

Guideline 5.4.1: Software Accessibility. If micro data are released as an information product, all micro-data products and documentation should be accessible without requiring the use of any single commercial product. Open-source formats (comma delimited, space delimited, etc.) must be available in addition to any others.

Guideline 5.4.2: File Description. Provide complete documentation for all data files.

- Data producers should determine what metadata standards are current at the time data files are prepared and produce associated metadata for their files that comply with applicable standards.
- Documentation must describe the data files including the title, data collection sources, tables that make up the set, inter-relation among tables (e.g., keys), and record layouts for data files.
- Documentation must also include descriptions for each variable in the data set that includes the variable name, description, type (categorical, numerical, date/time, etc.), format, entry restrictions (e.g., categories, range), and missing value codes.
- Indicate changes made to previously released data and the “as of” date of the data file.
- Provide a contact point with the data to facilitate user comments and suggestions.

Guideline 5.4.3: Information Quality Discussion. Micro-data files must discuss how the data were collected and the limitations of the data.

Guideline 5.4.4: Items Needed for Variance Estimation. Datasets containing sample data must contain weights and associated variables for accurate variance estimation. A dataset that requires weights and additional variables for the computation of estimates and standard errors should not be released before these items become available.

5.5 Rounding

Use consistent practices for rounding and displaying numbers in text, tables, and figures.

GUIDELINES

Guideline 5.5.1: Using Rounded Numbers. Make all calculations before rounding. Perform tabulations to produce summary data and computations for estimating standard errors on data as collected.

- The sum of the rounded numbers may not equal the rounded sum. In such a case, add a note indicating that totals may not equal the sum of their individual components due to independent rounding.
- To allow users to make further calculations accurately, do not round estimates disseminated in a spreadsheet.

Guideline 5.5.2: Degree of Rounding in Text and Graphs. The degree of rounding for text discussion and graphs should depend on the type of data (actual measure vs. sample), the known or suspected accuracy of the data, and the differences being discussed.

- Round percentages appearing in text to whole numbers unless smaller differences being discussed require decimal places and the accuracy supports it.
- Perform rounding consistently for similar subjects throughout the information product.
- In multiplying or dividing numbers, the resulting precision cannot be more precise than the least precise of the component numbers.

Guideline 5.5.3: General Rounding Rule. The general rules for rounding are:

- If the first digit to be dropped is less than 5, then do not change the last retained digit (e.g., round 6.1273 to 6.127).
- If the first digit to be dropped is 5 or greater, then increase the last retained digit by 1 (e.g., round 6.6888 to 6.69).

5.6 Data Protection

Release all information products in accordance with applicable Federal law and regulations along with any confidentiality pledges given to data providers.

GUIDELINES

Guideline 5.6.1: Non-disclosure of Confidential Data. For information collected under a confidentiality pledge, employ statistical disclosure limitation procedures and methods to protect any identifiable or other confidential data from disclosure prior to public dissemination. BTS staff must follow the established confidentiality procedures outlined in the latest agency confidentiality guidelines such as the BTS *Confidentiality Procedures Manual*.

Guideline 5.6.2: Security of Disclosure Limitation Methods. The BTS confidentiality officer must review and approve any descriptions of disclosure limitation methods prior to their public dissemination.

- Do not publish the details about how disclosure limitation methods were used to protect the data if publication could jeopardize data confidentiality. For example, do not reveal information on how noise may have been added to the data, what variables were used to implement record swapping, or the parameter values used to protect tabular data.

Guideline 5.6.3: Disclosure Review Requirements. All information products must be reviewed for compliance with the disclosure protection procedures stated in the latest BTS *Confidentiality Procedures Manual*.

5.7 Revisions

Create a standard process for handling possible post-dissemination data changes and document it.

GUIDELINES

Guideline 5.7.1: Scheduled Revisions. When appropriate, establish a schedule for anticipated revisions and make it available to users.

- Identify the first dissemination of a data value in an information product as "preliminary" if revisions are anticipated in a subsequent dissemination.
- Designate scheduled revisions to data values as "revised" (or “final”) when disseminating the changes.

Guideline 5.7.2: Errors in Previously Disseminated Information. Actions taken when data errors are discovered, or an external data source makes changes, depend on the impact that the potential revision would have on previously disseminated estimates.

- Establish threshold criteria for making revisions. For example, the threshold criteria might be to revise for changes exceeding five percent in smaller values or exceeding one percent in larger values.
- If the change does not exceed the threshold criteria, or threshold criteria do not exist, then management will determine whether the error is serious enough to warrant a revision.
- Document the error discovery and correction process.

Guideline 5.7.3: Documentation of Error Corrections. Document the nature of the changes, any corrective action needed to fix an error, and provide this information to data users.

- Identify data values changed due to unscheduled revisions and explain the reasons for these changes to data users.
- Document problems regardless of the scope of the error or the decision whether to revise the data.
- Provide error documentation to data users. Ensure timely and wide dissemination of information product revisions.

Guideline 5.7.4: Monitoring Revisions to Disseminated Data. Track the differences between an initial release of estimates and the corresponding final disseminated estimates for key data series.

- Examine the effect of revisions (number of times data are revised and the magnitude of the revisions). Revision error information can help users better understand the variability between initial estimates and final estimates. For data systems that are continuously updated, compare the initial estimates with estimates obtained after a suitable period has elapsed.
    - Some ways to present revision error information include the average revision error, the maximum revision error, or the distribution of revision errors during a specified time.

- If revision error for a key data series shows an initial release is an unreliable indicator of the final estimate, consider whether publishing the estimate with a measure of revision error or withholding the initial estimate is the best way to serve data users.

5.8 Public Documentation

Documentation for the public must include the materials and tools (if applicable) necessary to properly interpret and evaluate disseminated information.

GUIDELINES

Guideline 5.8.1: Source and Accuracy Information. Source and accuracy information should provide summary information suitable for posting on the web and should be regularly updated to include methodological changes and the results of any quality assessment studies. Source and accuracy statements should summarize:

- Data system objectives and frequency of information release,
- Target population and coverage, geographic or other characteristic distribution and, where applicable, sample selection methodology and sample size,
- Data collection methodology and content of unfilled forms,
- Data adjustments for missing data, nonresponse, coverage error, measurement error, seasonality, and (if applicable) confidentiality protection,
- Estimation methodology, including variance estimation methodologies for statistical samples,
- Description of major sources of error, including coverage of the target population, missing data effects, and measurement error, and
- A BTS point of contact for further questions and comments.

Guideline 5.8.2: Data Dictionary. Provide a data dictionary that defines data items in the data file, provides a high-level summary of data collection procedures, and documents appropriate uses for the data.

Guideline 5.8.2: Availability of Additional Documentation. To ensure the transparency of BTS information products, additional documentation (as specified in Chapter 2, Section 3.3, Section 4.6, Chapter 5, and Guideline 7.1.4) should be made available to customers upon request, unless such release would jeopardize confidentiality or disclose the actual methods used to protect the data.

Guideline 5.8.3: Reproducibility. Data users should be able to reproduce any publicly released information product to a reasonable degree of closeness. Information products that have been revised should clearly indicate the “as of” date.

Guideline 5.8.4: Archive Requirements. To ensure reproducibility within BTS, the product manager should establish criteria for retaining and archiving:

- All electronic product files,
- Complete information products, whether paper or electronic, representing a specific continuing publication product or one-time report,
- The data files and/or databases (at the most disaggregated level), which are used to generate publicly released information products, and

- System and model documentation and computer software/programs used to generate any information product.

Chapter 6 - Evaluating Information Quality

Ensuring data quality requires regular assessments of all aspects of data collection and processing and the implementation of corrective actions as appropriate. This can be accomplished by incorporating data quality checks within routine data collection processes and information product releases (Section 6.1), independently reviewing data products and data collection systems for standards compliance (Section 6.2), targeting evaluations to diagnose and resolve serious data problems (Section 6.3), and maintaining the frame (Section 6.4).

6.1 Data Quality Assessments

Statistical information products and the processes that agencies uses to create them should routinely include an evaluation component.

GUIDELINES

Guideline 6.1.1: Quality Assurance. All DOT information products and the processes that DOT uses to create them must routinely include:

- Process checks throughout data collection, data processing, and information dissemination,
- Pre-dissemination review of information products, and
- Measurement of performance and of information quality.

Guideline 6.1.2: Periodic Quality and Performance Self-Assessment. Product and project managers should periodically conduct a self-assessment of the quality and performance of their products and processes.

- Assess quality and performance on an annual basis for products and processes that occur at least once a year. Assess less frequently for products and processes that occur less than once a year.
- The assessment should highlight significant events that occurred during the past period, any events anticipated to occur during the next period, and identify strengths, weaknesses, and improvement opportunities.
- Submit a summary of the self-assessment findings to the Director or the Director's designated manager.
- Use assessment results to update internal and public documentation and to improve data quality.

6.2 Quality Control Processes

Independent statistical reviews enhance the credibility and effectiveness of data collection systems (including those handling external-source data) that are used to produce information products.

GUIDELINES

Guideline 6.2.1: Independent Review Team. An independent data quality review team should include:

- The senior agency official responsible for statistical methods and standards, who should establish the team,

- At least one person knowledgeable about statistical standards but not involved in the data collection process, and
- At least one person familiar with the data collection process.

Guideline 6.2.2: Review Areas. Independent data quality review should focus on compliance with statistical standards and with design specifications. The review should include:

- The most recent self-assessment report,
- The data collection design specifications,
- A historical review of problems identified by staff in collecting the data, the primary data users in applying the data to their needs, and data providers in reporting the data,
- If sampling is used, a review of the sample design and the sample selection and maintenance processes,
- A review of data processing problems, such as problems in converting raw data files to databases, problems with lack of editing or edit resolution, and problems with missing data,
- Review of the procedures for dissemination of data through various media, and of the source and accuracy information provided to users,
- Verification that users can independently reproduce estimates, including sampling error estimates where applicable, contained in the information products coming out of the system, and
- Verification that documentation is accurate, complete, and current.

Guideline 6.2.3: Review Outputs. Outputs from the data quality review should include:

- A report to the Director on review findings,
- A reply from the office responsible for the data collection system, and
- A quality improvement plan, prepared by the office responsible for the data collection system.

Guideline 6.2.4: Follow-up Review. A follow-up review should verify that the

improvements have been implemented.

6.3 Evaluation Projects

Agencies should undertake a data evaluation project if analysis of the data reveals that key data elements fail to meet data quality requirements.

GUIDELINES

Guideline 6.3.1: Evaluation Project Teams. An evaluation project team should report to the Director or to a designee with authority to allocate resources in support of the team's mission. The team members should include:

- A team leader who is not involved in the data collection process,
- Personnel selected for their expertise but not involved in the data collection process, possibly including non-BTS staff, and
- Personnel who are directly involved in the data collection process.

Guideline 6.3.2: Evaluation Plan. The project team should plan the evaluation as a type of data analysis that targets specific problems or issues in BTS data products. Solicit input from the following sources, with the greatest weight given to the primary users:

- Primary users for whom the agency designs information products, and normally include analysts in DOT and the Congress,
- Secondary users, including commercial interests and the general public,
- Data collection experts who can identify additional process quality issues, and
- Sponsored independent expert reviews.

Guideline 6.3.3: Conduct of an Evaluation Study. The major tasks in an evaluation study are:

- Specifying the processes contributing to the observed data quality problem,
- Identifying the root problems,
- Ascertaining solutions to the problems, and
- Drafting an improvement plan to address the problems identified. In some cases, the recommendations could lead to a redesign of the data collection system (Chapter 2).

Guideline 6.3.4: Implementation of the Study Recommendations. The project team should report the evaluation results and recommendations to the senior agency official responsible for statistical products. Upon concurrence, the office conducting the data collection is responsible for the implementation of the recommendations. A follow-up data evaluation should verify that the recommendations have been implemented and that the data meet quality requirements.

6.4 Frame Maintenance and Updates

Frames (lists of potential data providers) must be maintained, updated,

evaluated, and archived to ensure that coverage is as complete and current as possible. =

GUIDELINES

Guideline 6.4.1: Maintaining Coverage. Frames must be maintained and updated.

- Maintenance is the continuous revision of the frame based on new information that becomes available during data collection. For regulatory or administrative data collections, frame maintenance requires that changes related to reporting eligibility are promptly reflected in the data collection system.
- Updates are systematic, comprehensive searches for frame changes that canvass all available information. Updates can also include re-examination of reporting categories using more recent information, such as reclassifying airlines based on annual operating revenues. Maintenance and updating actions include:
- Additions of new potential data providers,
- Revisions due to changes in ownership, name, or address.
- Changes in how data providers are classified (for reporting or sampling purposes), and
- Deletions of data providers no longer in the target population.

Guideline 6.4.2: Coverage Evaluation. In addition to routine maintenance and updates, periodically evaluate target population coverage of frames that are used for recurring data collections.

- The frequency of coverage evaluations depends on the relative stability of the target population and on the frequency of data collection.
- Evaluate coverage of administrative or regulatory data collections at least annually.
- If the frame is properly maintained and updated, problems in coverage for regulatory based systems can be avoided.
- Conduct an evaluation of the potential bias if the frame's coverage of the target population falls below 85 percent.

Guideline 6.4.3: Archiving. Frames are a critical component of data collection and documentation. A backup copy of the current frame must be created and archived prior to each major frame update (or periodically, for continuously maintained frames).

- All active and inactive data providers must be included on the archive file.
- Inactive records may be periodically deleted from the current file, after the prior file has been archived.
- During a frame update, information on potential data providers should not be deleted from the frame. Instead, a status indicator field in the frame should designate whether the entry is active/inactive or in-scope/out-of-scope.
- Whenever the information contained in a frame is modified, record the effective date of the change.
- Provide a way of tracking changes in frame record identifiers over time.

Guideline 6.4.4: Frame Maintenance Documentation. Documentation for maintaining and updating frames must be written and revised as necessary. The documentation must include:

- The frequency of routine maintenance and major updates,
- Sources of information used for maintenance and updates,
- Procedures for incorporating the results of the updates on all appropriate files, mailing lists, and other data collection control forms or listings
- Summary of results of the frame maintenance and updates, and
- The results of periodic coverage studies.