Transparency in the Reporting of Quality for Integrated Data: International Standards

Third FCSM/WSS Workshop on Quality of Blended Data
Reporting on Quality Issues in Output Data

February 26, 2018

John L. Czajka
Mathew Stange

What international standards may offer

- **Administrative data systems more developed**
- **Decline in survey response rates—at least in Europe—more rapid**
- **International organizations have been particularly active in development of standards**
  - **Eurostat and the European Statistical System; United Nations**
  - **Recent focus on use of administrative records and Big Data for official statistics**

European Union statistical organizations

- **Eurostat is a Directorate General of the European Commission, the executive of the European Union**
  - **Eurostat is the statistical office of the European Union**
  - **Eurostat is charged with the production of official statistics at the level of all Europe for the European Union**
- **European Statistical System (ESS) is a partnership between Eurostat and the statistical authorities of the member states**
  - **ESS Committee charged with providing "professional guidance to the ESS for developing, producing, and disseminating European statistics"**

Key documents from European Union

- **European Statistics Code of Practice for the National and Community Statistical Authorities (2011)**
- **Quality Assurance Framework for the European Statistical System (2015)**
- **ESS Handbook for Quality Reports (2015)**
  - **Includes in an appendix:**
    - ESS Guidelines for the Implementation of the ESS Quality and Performance Indicators

European Statistics Code of Practice

- **Delineates 15 principles that address:**
  - **The institutional environment (principles 1 – 6)**
  - **Statistical processes (principles 7 – 10)**
  - **Statistical output (principles 11 – 15)**
- **With some variation, the principles for statistical output are often included in frameworks of individual countries inside and outside Europe; often described as "dimensions of quality"**

15 principles of Code of Practice

- **Institutional environment**
  1. Professional independence
  2. Mandate for data collection
  3. Adequacy of resources
  4. Commitment to quality
  5. Statistical confidentiality
  6. Impartiality and objectivity
- **Statistical processes**
  7. Sound methodology
  8. Appropriate statistical procedures
  9. Non-excessive burden on respondents
  10. Cost effectiveness
- **Statistical output**
  11. Relevance
  12. Accuracy and reliability
  13. Timeliness and punctuality
  14. Coherence and comparability
  15. Accessibility and clarity

Indicators of compliance

- **For each principle, the Code lists several indicators, which describe actions that conform to the principle**
- **For example, under accuracy and reliability:**
  - **12.1 Source data, intermediate results and statistical outputs are regularly assessed and validated**
  - **12.2 Sampling errors and non-sampling errors are measured and systematically documented according to the European standards**
  - **12.3 Revisions are regularly analyzed in order to improve statistical processes**
- **The Code does not discuss these indicators; that is left to the ESS Quality Assurance Framework**

ESS Quality Assurance Framework

- **Framework produced to assist national statistical organizations in implementing the Code of Practice**
- **Designed as an aid in achieving quality—not measuring or reporting it**
- **Provides series of methods at both the institutional and product/process levels to facilitate achievement of the goal expressed in an indicator**
  - **For example, at the product/process level three methods for indicator 12.2:**
    - Periodic quality reporting on accuracy is in place
    - Quality reporting on accuracy is guided by ESS recommendations
    - Methods and tools for preventing and reducing sampling and non-sampling errors are in place

ESS Handbook for Quality Reports

- **Purpose is "to provide guidelines for the preparation of comprehensive quality reports for a full range of statistical processes and their outputs"**
- **Specific objectives of these guidelines:**
  - **To promote harmonized quality reporting across statistical processes and their outputs within a Member State and hence to facilitate comparisons across processes and outputs**
  - **To promote harmonized quality reporting for similar statistical processes and outputs across Member States and hence to facilitate comparisons across countries**
  - **To ensure that reports include all the information required to facilitate identification of statistical process and output quality problems and potential improvements**

Quality reports and quality profiles

- **Comprehensive quality reports addressed by the Handbook bear resemblance to U.S. quality profiles**
- **A survey quality profile summarizes what is known about the sources and magnitudes of errors in a survey (Kasprzyk and Kalton 2001)**
  - **A systematic and comprehensive review across the spectrum of survey activities in which both qualitative and quantitative results are brought together to allow an assessment of the quality of the survey operations and the data**
  - **Relevance, timeliness, and accessibility are dimensions of quality not usually treated in quality profiles in the U.S.**
- **Quality profiles were produced for several federal surveys**
- **They are resource intensive, they require information that may not exist, and their value to the survey producer is questionable**

Handbook guidelines

- **Handbook provides guidelines specific to each of the five dimensions of statistical output quality plus three others:**
  - **Confidentiality (principle 5)**
  - **Burden (principle 9)**
  - **Cost (principle 10)**
- **Handbook also includes guidelines on statistical processing, which is not one of the Code principles**
- **Recommendations for quality reporting include 16 quantitative indicators for the five quality dimensions**

Integrated data in the European framework

- **ESS Quality Assurance Framework and Handbook do not purport to be directed at integrated data but acknowledge that some of the estimates produced by European nations may be based on integrated data**
- **Under the accuracy dimension there are separate discussions of statistical processes using administrative sources and statistical processes involving multiple data sources**
- **Also a general recommendation that whenever multiple data sources were used, a separate quality report should be produced for each data source and not just the combination of multiple data sources**

Processes using administrative sources

- **Over- and undercoverage loom large as error sources**
- **Other error sources include:**
  - **Non-response at the unit and item levels**
  - **Measurement error**
  - **Processing errors by the provider or statistical agency**
  - **Conceptual differences between the register and the target**
- **Multiple registers necessarily involve some linkage, the quality of which depends in part on the quality of the identifiers**

Processes using multiple data sources

- **When processes involve multiple data sources, the individual components should be assessed, but focusing on the "whole picture" is necessary as well**
- **A quality report should include how the process is organized, the individual segments that are included, and a summary of the quality aspects**
- **The only suggestion regarding assessment of the quality of the final product applies only when a preliminary estimate is followed by a revision**
  - **The magnitude of the revision may be indicative of quality**
  - **Small revisions suggest conversion on the true value**
  - **Yet overall revisions may not address all sources of error in the initial estimates**

Other quality dimensions

- **Relevance**
  - **Focus is on users of the statistical outputs and to what extent the data satisfy their needs**
  - **Different groups of users may have different needs**
  - **The one quality and performance indicator is the data completeness rate: the ratio of data cells provided to cells required**
- **Timeliness and punctuality**
  - **Quality and performance indicators include:**
    - Time lag between end of reference period and initial results
    - Time lag between end of reference period and final results
    - Time lag between delivery of data and announced target date

Other quality dimensions cont'd

- **Coherence and comparability**
  - **This dimension is assigned high importance, with extensive information requested for the quality report**
  - **Quality and performance indicators address only "mirror flows" (inflows and outflows that should match) and length of unbroken time series**
  - **A caution not to confound coherence/comparability with accuracy (seeming inconsistency could be due to inaccuracy)**
- **Accessibility and clarity**
  - **User feedback is the best source of information in addressing this dimension in the quality report**
  - **What can more sophisticated and less sophisticated users access?**
  - **Quality indicators include how often users consult tables and metadata and the degree of completeness of the latter**

Other principles

- **Cost**
  - **Quality report should include cost breakdown by major components although difficulty of obtaining this is noted**
- **Burden**
  - **Quality report should include:**
    - Respondent burden in financial terms or hours
    - Targets for reducing burden and recent efforts to reduce burden
    - Whether information collected is limited to what is absolutely necessary and cannot be obtained elsewhere
- **Confidentiality**
  - **Distinction between legal requirements and data treatment**
  - **Not mentioned are measures to assess effectiveness**

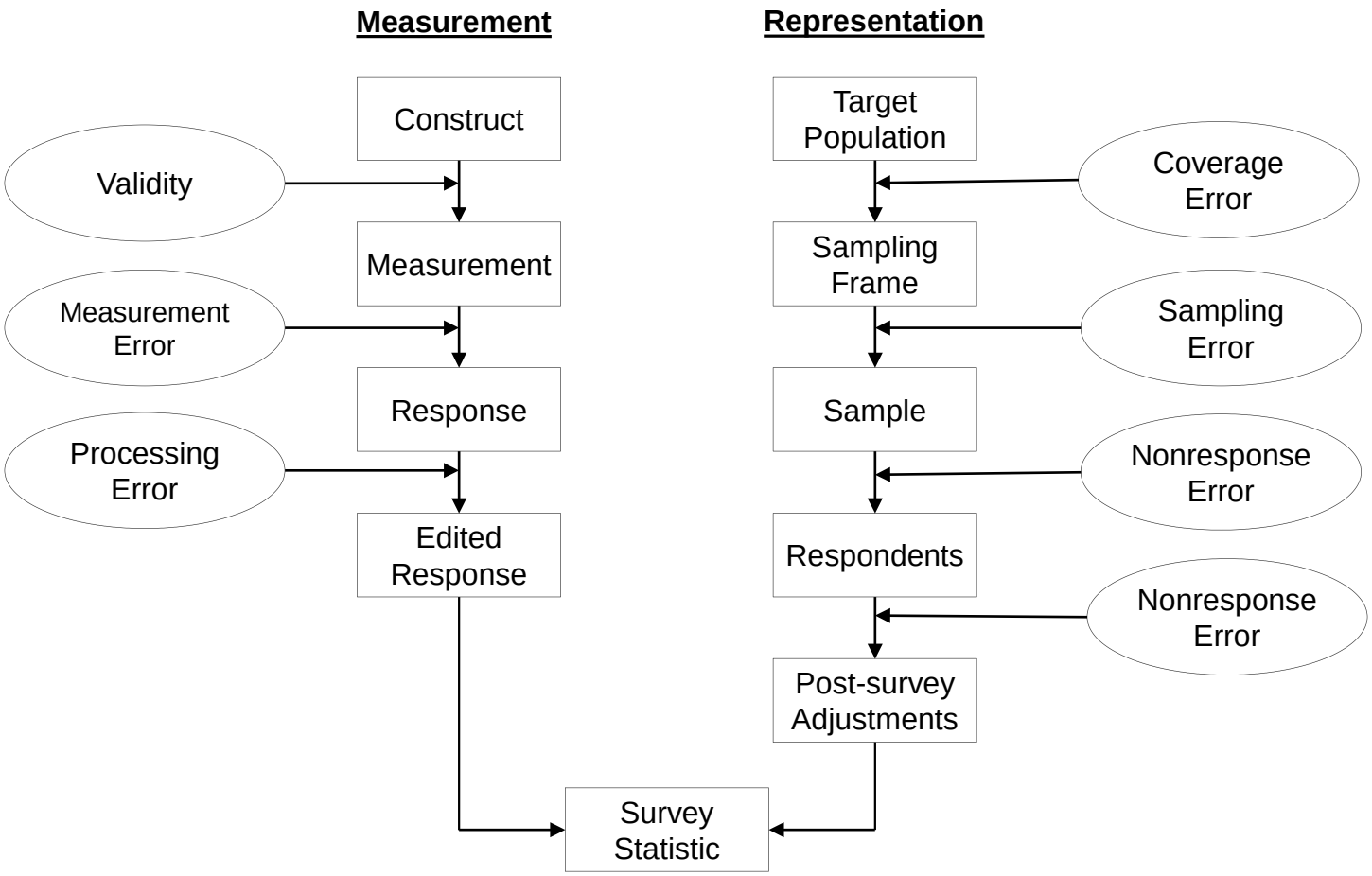
Extending TSE to integrated data

- **Li-Chun Zhang of Statistics Norway has proposed a framework for integrated data based on the life cycle model of Total Survey Error (TSE) in Groves et al. (2009)**
- **Statistics New Zealand has adopted this framework as the basis for its own quality framework for integrated data**

Overview of the TSE life cycle model

- **The TSE model follows the life cycle of a survey from conception to the production of a survey statistic**
- **The model builds on the idea that a sample survey consists of questions administered to a sample drawn from a target population**
- **The model traces the dimensions of measurement and representation from an abstract construct and a target population through the design and implementation of a survey, culminating in a survey statistic**
- **Error may be introduced at each stage as depicted in the figure on the next slide**

Survey life cycle from quality perspective

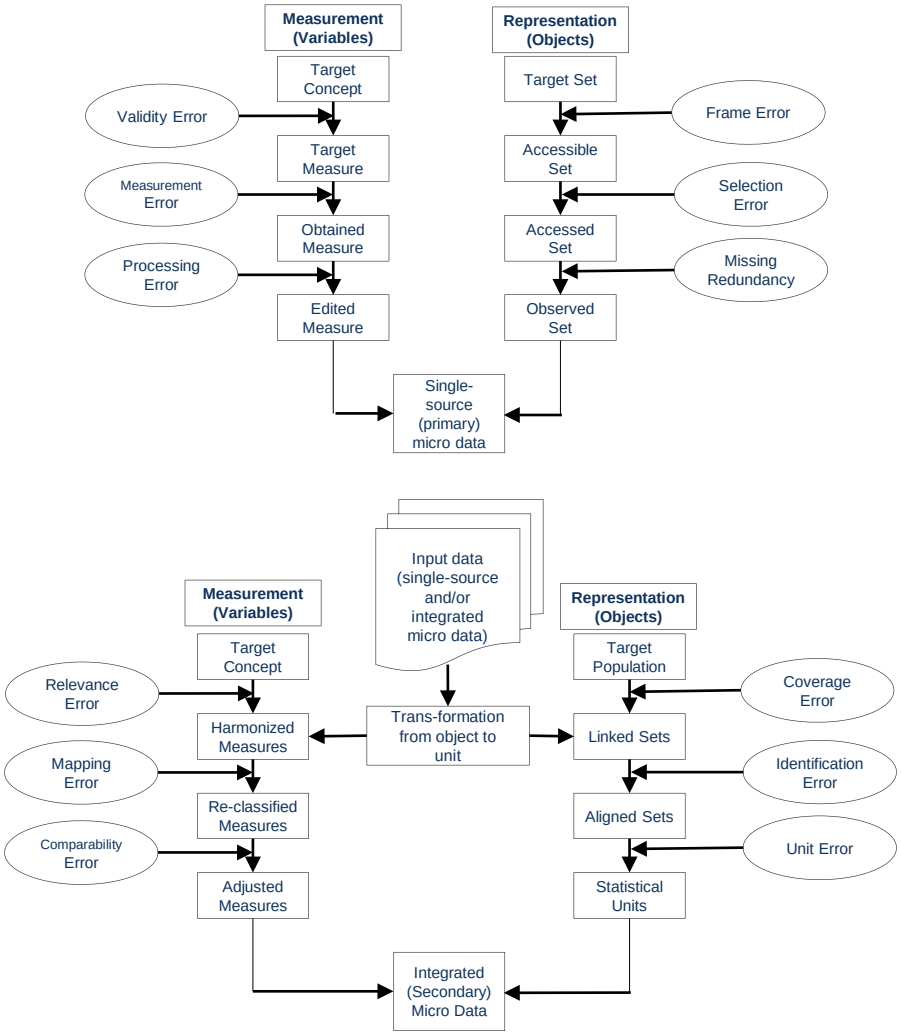


Source: Groves et al. (2009).

Two-phase life-cycle model

- **In Zhang's two phase model, the end result of each phase is a micro dataset—not a single statistic**
- **In addition, most of the concepts from Groves et al. have been renamed to accommodate the inclusion of data from administrative sources**
  - **For example, "measures" in place of "responses" and "sets" in place of "sample" and "respondents"**
- **Phase one describes a single microdata source, but each input to the integrated microdata has its own phase one assessment**
- **Phase two shows the multiple inputs and depicts the sources of error for the integrated microdata**

Two-phase life cycle of integrated data



Source: Zhang (2012).

Elements of the two-phase model

- **Zhang observes: "the 20th century witnessed the birth and maturing of sample surveys; the 21st century will be the age of data integration"**
- **Harmonization on the measurement side and linkage on the representation side are steps in phase two**
- **On the representation side, Zhang uses "objects" in phase one and "units" in phase two; the trans-formation of objects into units is shown in a box in phase two below the input of multiple data sets**
  - **Phase one data may include, for example, jobs while the goal of integration may be data on persons**
  - **Units themselves may have to be combined in some way—for example, persons aggregated to households**

Error at the dataset level

- **Zhang's conceptualization envisions an ideal target integrated dataset—the analog to an error-free survey statistic**
- **Discrepancies between the target dataset and the final integrated dataset are analogous to the concept of total survey error in Groves et al.**
- **To assess the accuracy of the final dataset, Zhang develops the concept of empirical equivalence**
  - **Two datasets are empirically equivalent if they generate identical inferences; this does not require micro-level equivalence**
- **Zhang extends empirical equivalence to the assessment of public use data, where error is introduced to protect confidentiality**

Statistics New Zealand (Stats NZ)

- **With a mandate to make administrative data the data source of choice, Stats NZ faces the need to "assess and explain the quality of statistics that use multiple sources, including administrative data" (Holmberg and Bycroft 2017)**
- **Stats NZ issued a *Guide to Reporting on Administrative Data Quality*, which uses Zhang's framework**
  - **Includes quality indicators for each of the phase one and phase two error sources**
  - **25 quantitative indicators for phase one and 19 for phase two**
  - **34 qualitative indicators for phase one—mostly descriptive**
  - **No qualitative indicators as yet for phase two**

Quantitative indicators for phase two

Representation dimension

Coverage error

1 Undercoverage
2 Overcoverage
3 Proportion of units linked from each dataset to a base dataset, or percentage link rates between pairs of datasets
4 Proportion of duplicated records in the linked data
5 Precision and recall in linking
6 Macro-level comparisons of the distribution of linked objects with reference distributions
7 Delay in reporting
8 Linking methodology used

Identification error

9 Proportion of units with conflicting information
10 Proportion of units with mixed or predominance-based classifications
11 Rates of unit change from period to period

Unit error

12 Proportion of units that may belong to more than one composite unit

Measurement dimension

Relevance error

13 Percentage of items that deviate from Statistics NZ/international standards or definitions

Mapping error

14 Proportion of items that require reclassification or mapping
15 Proportion of units that cannot be clearly classified or mapped
16 Distribution of variables in linked data
17 Indicators and measures of modeling error

Comparability error

18 Proportion of units failing edit checks
19 Proportion of units with imputed values

Adding a third phase to the framework

- **Reid et al. (2017) added a third phase for assessing the quality of final outputs—that is, the statistical estimates derived from the integrated microdata that is the endpoint of phase two**
- **Quality indicators do not yet exist for phase three**
- **Reid et al. provide three case studies that illustrate different approaches to evaluation**
  - **Case study 1**
  - **Case study 2**
  - **Case study 3**

Administrative data and official statistics

- **Daas et al. (2011) present quality indicators for administrative data used as an input to official statistics**
- **Indicators address five dimensions:**
  - Technical checks
  - Accuracy
  - Completeness
  - Integrability
  - Time-related factors
- **Indicators draw on phase one of Zhang (2012) in corresponding to objects (representation) versus variables (measurement)**

Indicators of integrability

- **Dimension of integrability bears most directly on the integration of multiple sources**
- **Four indicators are intended to capture how well the data source can be integrated into the statistical production system of an organization**
  - **Objects**
    - Similarity of objects in source with those used by organization
    - Ability to align objects in source with those of organization
  - **Variables**
    - Usefulness of linking variables in source
    - Closeness of variables in source with those in other sources used by the organization

Big Data for official statistics

- **The U.N. created a Global Working Group on Big Data, which is working toward standards**
- **Multiple teams are addressing different aspects**
- **Big Data Quality Task Team published *A Suggested Framework for the Quality of Big Data* (2014)**
  - "The application of either traditional data quality frameworks or those designed for administrative data would be an inadequate response to Big Data"
- **Broader scope of Big Data compared to administrative data required a different approach**

Big Data quality framework

- **11 dimensions are nested within the hyper-dimensions of Source, Metadata, and Data and applied to the phases of input and output although not to the broad, throughput phase**
- **Possible indicators are listed for each dimension**
- **Most indicators posed as questions although some specify calculations**
  - For example, for the input phase assess coverage error, duplicates, representation of sub-populations, and calculate an R-index
  - Indicators for the output phase are considered less useful
- **Framework is in early stages of development—clearly a work in progress**

For More Information

- **John L. Czajka**
  - **JCzajka@mathematica-mpr.com**

- **Mathew Stange**
  - **MStange@mathematica-mpr.com**