# Overview of Transparent Quality Reporting in the Integration of Multiple Data Sources

**John L. Eltinge**
**Assistant Director for Research and Methodology**
**U.S. Census Bureau**
**Chair, Federal Committee on Statistical Methodology**

**FCSM Conference Session A-5**
**March 7, 2018**

# Acknowledgements and Disclaimer

This presentation summarizes some elements of work by dozens of colleagues on the FCSM Working Group on Data Quality, the Committee on National Statistics, and participants in related workshops and meetings. Any errors are the responsibility of the presenter.

The views expressed here are those of the presenter and do not necessarily represent the policies of the United States Census Bureau.

# Overview: Transparent Quality Reporting

I. Context and Goals

II. Work to Date

III. Initial Report

# I. Context and Goals

A. Historical focus of statistical agencies:

Use sample surveys (with some other sources) to produce high-quality statistical series, some public-use microdata

# I. Context and Goals (continued)

B. Changing environment:

1. Declining survey response rates, increasing costs, increasing expectations of data users

2. Increasing availability of multiple data sources (beyond surveys)
   Ex: admin, commercial, sensors

# I. Context and Goals (continued)

C. Opportunity: Integrate multiple data sources to:

1. Improve the balance of quality, risk and cost for current statistical production

2. Expand the suite of statistical information products and services in priority areas (geography, time, refined models)

# I. Context and Goals (continued)

D. Starting Point:

Transparent Reporting in High-Priority Areas of:

1. Quality: Accuracy, timeliness, relevance, comparability, coherence, accessibility

2. Risk: Production failures, disclosure

3. Cost: Cash, scarce skills, respondent burden

# I. Context and Goals (continued)

E. Emphasize Distinction Between:

1. Now - Transparent Reporting: What We Do/Know?

   Ex: AAPOR stds for computing response rates

2. Later - Specific Numerical or Operational Standards

   Ex: Response rate must be at least X%

   Not yet for integration of multiple sources,

   until informed by trajectory of experience

***Columns: Performance Dimensions***

| ***Rows: Areas for standards*** | **Quality (accuracy)** | **Quality (other dim)** | **Risk** | **Cost** |
|---|---|---|---|---|
| **Transparent reports for users** | ***Current emphasis*** | ***Additional discussion*** | | |
| **Transparent rep to improve** | ***Additional discussion*** | ***Additional discussion*** | | |
| **Research, design production, empirical results** | | | | |
| **Legal, regulatory privacy areas** | | | | |

# II. Work to Date

A. Three public workshops (with the Washington Statistical Society)

Input data quality (12/1/2017)

Processing quality (1/25/2018)

Output data quality (2/26/2018)

Additional events planned

# II. Work to Date (continued)

B. Meetings with the Committee on National Statistics, other stakeholders: Identified

1. Well-developed quality frameworks (CNSTAT, ESS)

2. Related standards (often survey-centric) from OMB, agencies (U.S. and international), professional groups (e.g., ISO)

# II. Work to Date (continued)

B.3. "Quality profiles" - some U.S. stat programs

B.4. Central themes:

- "Fitness for use" – context/user-specific

- Communication with identified audience: general public, "power users," technical

# II. Work to Date (continued)

C. Media and the General Public

Imprimatur as "trustworthy"

i.e., trusted independent source can verify

- Open to independent external scrutiny?
- Follows predetermined procedures?
- Other general criteria as in *Principles and Practices of a Federal Statistical Agency* ?

# II. Work to Date (Continued)

D. “Power Users” of Specific Series

Input quality: Sources and limitations clearly stated; any “black box” issues identified

Processing quality: Follows reasonable and customary procedures?

# II. Work to Date (Continued)

D. “Power Users” of Specific Series (continued)

Output quality:

- Consistent w/other comparable information?

- Timely identification and explanation of major changes and inconsistencies?

# II. Work to Date (Continued)

E. Technical Specialists

Ideal (unattainable?): Sufficient information for

- Reproducibility of results (intermediate or final output)

or perhaps even

- Full replicability of all steps: unit capture, data collection, record linkage, imputation, …, final estimates

# II. Work to Date (Continued)

E. Technical Specialists (continued)

Realistic step: Sufficient information on

- Main steps of full production process

- Strengths and limitations in quality dimensions most likely to effect stakeholder value

# II. Work to Date (Continued)

E. Technical Specialists (continued) – examples:

Input: (Sub)population coverage rates, item missingness, variable definitions, timeliness

Models in processing: Outcome variable, predictors, functional form, outlier treatment, estimation methods, diagnostics

# II. Work to Date (Continued)

E. Technical Specialists (continued) – examples:

Output: Extending small domain practice

- Magnitudes of bias & variance terms?
- Prominent special cases – break in series?
- Sensitivity analyses?

Implementation: Systems well designed, documented, tested and maintained?

# III. Initial Report

A. Goals and overview

B. Elements of transparent reporting on selected quality dimensions

C. Applying elements to some current products

D. High-priority open questions in research and practice (Session H-5: 10:30 – 12:15 on March 9)

# Thanks to all

## Comments and questions welcome:

## John.L.Eltinge@census.gov

# Supplementary Questions

A. General Questions:

In using data products (especially based on integration of multiple data sources):

1. Predominant worries about quality?

# Supplementary Questions (Continued)

2. Impact of quality problems on practical value for your data users: Concrete cases

   a. How specific data series are used by your key stakeholders

   b. Specific quality issues that can degrade value of (a)?

# Supplementary Questions (continued)

2.c. Efforts you make to mitigate (b)?

2.d. How transparent reports on specific quality elements can help stakeholders understand (b), mitigate (c) and choose among competing data series?

2.e. Examples of good practice in (c) and (d)?

# Supplementary Questions (continued)

A.3. Best ways to communicate on (2) with non-specialists:

a. Criteria for "high quality data series"
   Ex: Checklist for "transparent reporting"
   Ex: Checklist (or longer reports) on specific quality features?

b. Why (a) is important for them?

# Supplementary Questions (continued)

B. Examples (conversation starters):

1. Break in series

a. Outright loss of data source

b. Changes in data capture and management systems

Ex: Duplication of records

# Supplementary Questions (continued)

1.c. Level shift (or changes in stability, seasonality) from (undetected?) changes in:

- (sub) population coverage

- accounting methods in administrative or commercial records

# Supplementary Questions (continued)

B.2. “Apples and oranges”

- Differences within or across data sources

a. Conceptual or operational definitions

Ex: “employment” – W-2? 1099? 1120S?

Ex: “sale” when ordered, delivered, paid?

b. “Unit” definitions: firm/establishment, geo

# Supplementary Questions (continued)

B.3. Relevance:

Ex: Administrative or commercial record systems may not keep up with true economic phenomena

B.4. Many other examples

# Thanks to all for your insights

# Additional comments welcome:
# John.L.Eltinge@census.gov