Quality Standards for Acquisition and Use of Multiple Data Sources

John L. Eltinge
U.S. Bureau of Labor Statistics

FCSM Policy Conference
Session #5

December 6, 2016

Acknowledgements and Disclaimer

The author thanks many colleagues in the Federal Committee on Statistical Methodology (FCSM), the FCSM Subcommittee on Administrative Records and the FCSM Subcommittee on Administrative, Alternative and Blended Data, BLS, the federal statistical system, academia and the private sector many productive discussions of these topics over the past two decades.

The views expressed here are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

Overview

I. Why Have Standards for Data Quality?

II. Dimensions Addressed by These Standards?

III. Impact of Standards on the Balance of Quality, Risk and Cost

I. Why Have Standards for Data Quality?

A. Mission of Government Statistical Agencies:

Provide public with high-quality information on a sustainable and cost-effective basis

B. Questions: Can data quality standards

1. Help statistical agencies fulfill their mission?
2. Give key stakeholders clear and accessible indications of the strengths and limitations of our statistical information products and services?

I. Why Standards? (Continued)

C. General Reasons for Standards in Technical Fields:

1. Improve value of product or service ("raise the bar")

2. Within the field:

   a. Common language, operating conditions

   b. Reduce transaction and integration costs

I. Why Standards? (Continued)

3. Customers:

   a. Broad: Have confidence in product or service (analogy: food safety)

   b. Practical distinctions on value delivered: Gold/silver/bronze level

   c. More refined: Accessible summary of what is (and is not) provided – (motor oil: 10 W 30)

II. Dimensions Addressed by Standards?

A. Broad & deep societal reconsideration of (statistical) information over decades:

- expectations on quality, cost, risk, credibility, accountability, access, stakeholder utility

- tools to address

- new products

- resource allocation: amount and mechanisms

II. Dimensions (Continued)

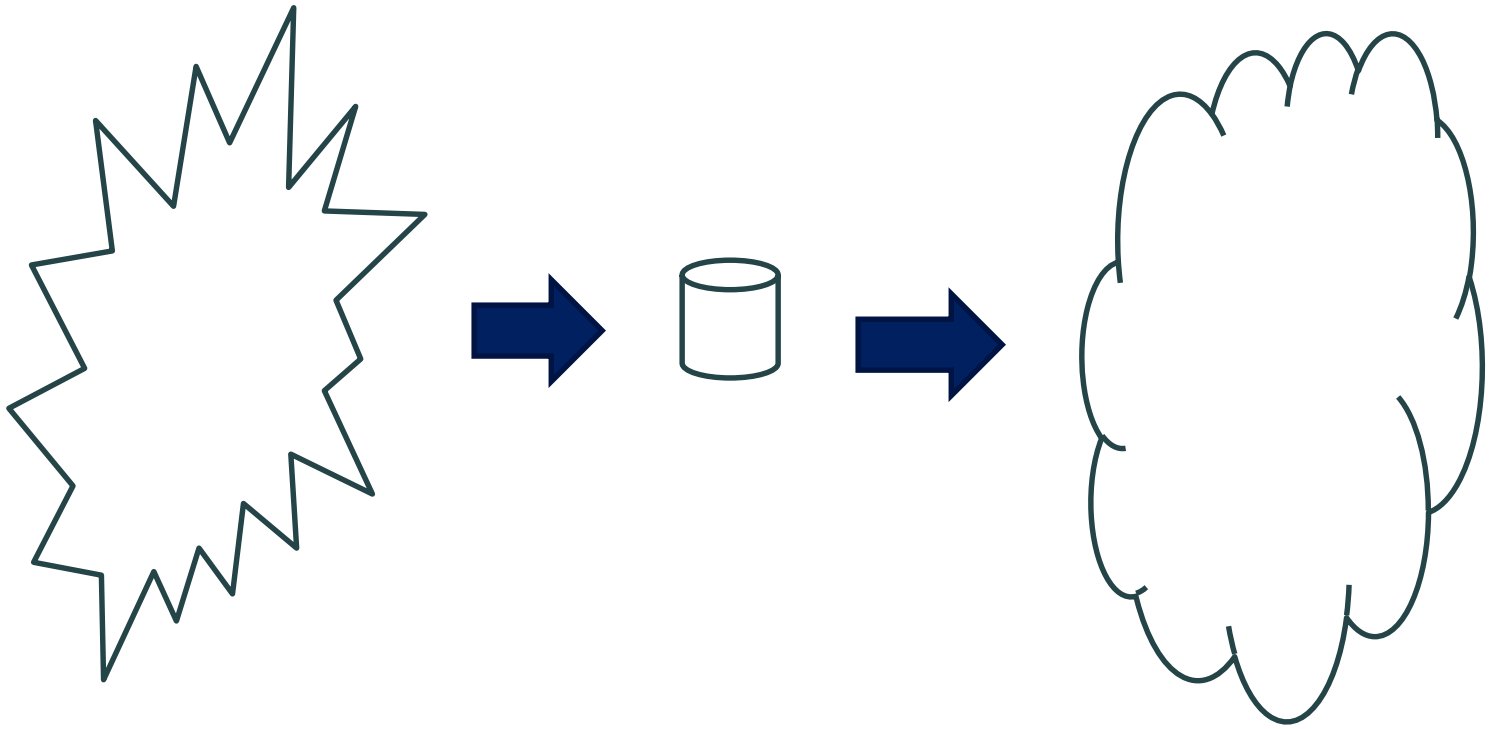
B. Increasing Importance of:

1. Wider range of inferential goals

2. Multiple data sources, beyond surveys:
   "Big data" (or "organic data" "non-designed data" or "alternative data" Groves, 2012; Couper, 2013)

3. Both (1) and (2) require thoroughgoing examination of previously developed standards

Standards for What? Inputs? Outputs? Processes?



Prospective Data Sources

Agency Processes

Prospective Information Products

II. Dimensions (Continued)

C. Output Data Quality - Multidimensional Definition

1. Example: Brackstone (1999): “fit for use”

   accuracy, relevance, timeliness, coherence, comparability, accessibility

   - others add transparency, other dimensions

2. Historical focus: Accuracy, usually through “total survey error” terms, reproducibility

II. Dimensions (Continued)

3. "Accuracy" component: "Total survey error" decomp

(Estimator) – (True value)

= (frame error/coverage); representativeness
+ (sampling error); can reduce with "big data"
+ (incomplete data effects); unit, period, item
+ (measurement error); unit, specification err
+ (processing effects); include model lack of fit

Large literature

II. Dimensions (Continued)

D. Input data quality

1. Align with impact on output quality
   - coverage, incomplete-data patterns, specification issues, measurement errors

2. Link with literature on quality management and standards for complex supply chains

II. Dimensions (Continued)

E. Statistical processes

1. Idealized: Statistical efficiency, conditional on specified set of input quality profiles and desired output quality profiles

2. Also consider robustness of processes against shocks

   - adaptation from engineering literature on "fault-tolerant designs"

II. Dimensions (Continued)

3. Management of Risks: Four Possible Outcomes

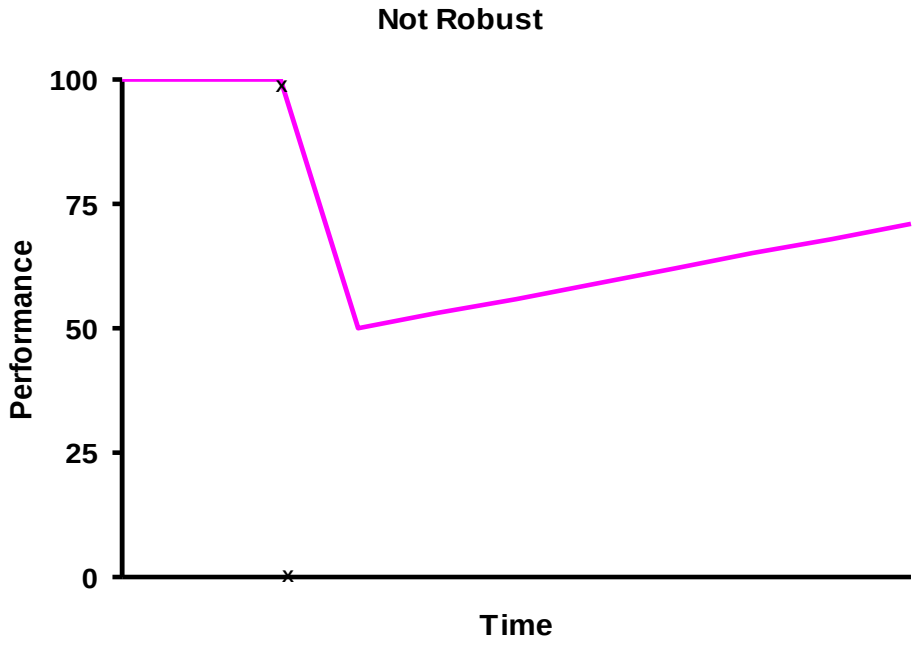
    1. Perfect resilience

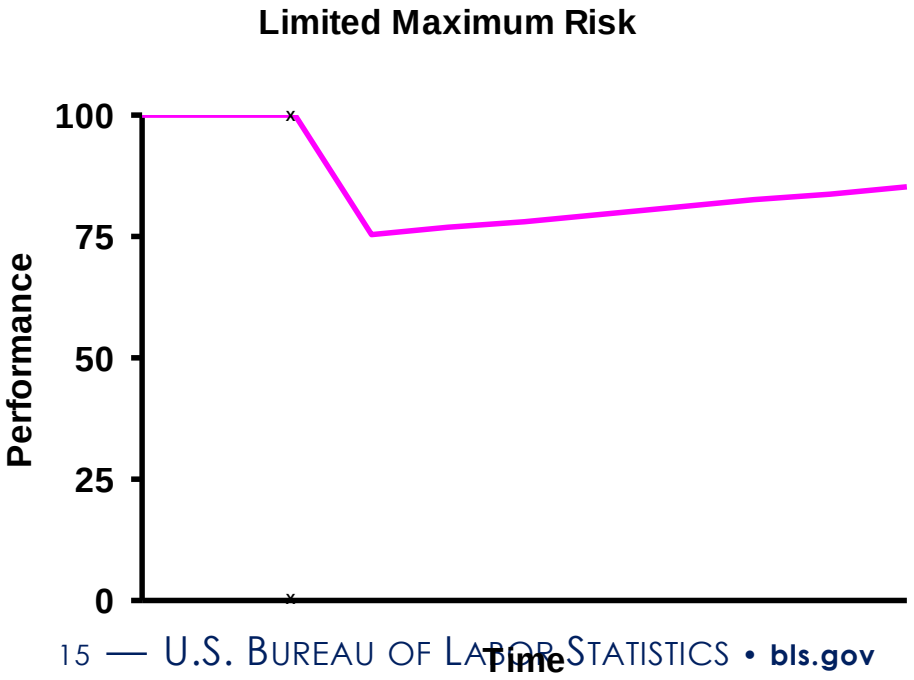
    2. Substantial degradation in quality with slow recovery

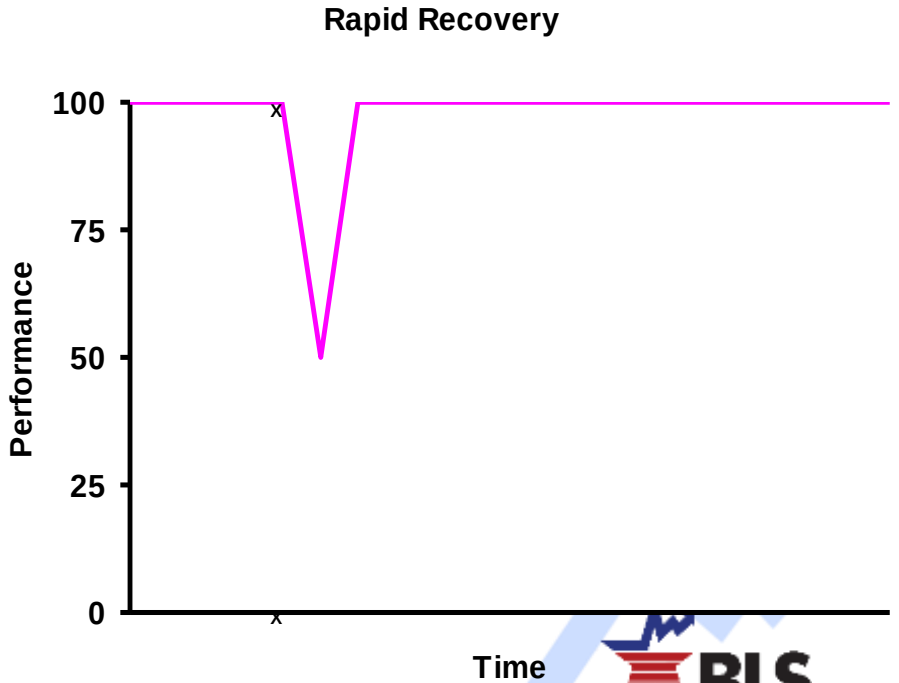
    3. Moderate degradation, again with slow recovery

    4. Substantial degradation with rapid recovery (cf. literature on “fault-tolerant designs”)


Perfect Resilience
100
75
50
25
0
Performance
Time


Not Robust
100
75
50
25
0
Performance
Time


Limited Maximum Risk
100
75
50
25
0
Performance
Time


Rapid Recovery
100
75
50
25
0
Performance
Time

III. Impact of Standards on Quality, Risk and Cost

A. Prospective role of standards for quality, risk profiles and cost structures:

1. Outcomes: Numerical criteria - CVs, linkage rates

2. Process: measurement, modeling & management, plus curation

3. Integrity, transparency and replicability, in forms that resonate with key stakeholders

4. Qualifications of personnel, organizations

III. Impact of Standards (Continued)

B. Prospective Limitations:

1. Technical development: Ready for standards?

2. Heterogeneous group of data users:
   Different standards applicable

3. "Minimum standard" can become "maximum quality"
   especially under resource constraints, unless absent clear incentives to meet higher standards ("platinum/gold/silver/bronze")

IV. Closing Remarks

A. Why Have Standards for Data Quality?

B. Dimensions for Standards: Output, Input, Processes

C. Impact of Standards on the Balance of Quality, Risk and Cost

Contact Information

John L. Eltinge
Associate Commissioner
Office of Survey Methods Research
202-691-7404
Eltinge.John@bls.gov