Data Quality Session

Developing a Data Quality Profile for the Consumer Expenditure Survey

Yezzi Angi Lee

Veri Crain, Scott Fricker , Evan Hubener, Clayton Knappenberger, Brandon Kopp, Julie Sullivan, and Lucilla Tan.

July 18, 2017

Presentation Outline

To share the challenges encountered in the initial stages of this development process, report on interim progress, and thoughts for next steps.

- ✓ What is a Data Quality Profile (DQP)
- ✓ Challenges
- ✓ Iterative approach to development
- ✓ Interim results
- ✓ Moving forward

What is a Data Quality Profile(DQP)?

"A comprehensive report prepared by producers of survey data that provided information data users need to assess the quality of the data"

Survey Research Center (2010)

" To provide researchers and data users with a single source for a wide range of information on the quality of AHS data"

Quality Profile of the American Housing Survey (1996)

More Example:
Vary in Breadth and Depth of Coverage

BRFSS 2013 Summary Data Quality Report

Table of Contents

- ✓ **RESPONSE RATES**
- ✓ **23 PAGE**
- ✓ **Annual publication**

https://www.cdc.gov/brfss/annual_data/2013/pdf/2013_dqr.pdf

American Housing Survey 1996 Quality Profile

- ✓ **TOTAL SURVEY ERROR DIMENSIONS**
- ✓ **80 + PAGE**
- ✓ **1996**

https://www.census.gov/content/dam/Census/programs-surveys/ahs/publications/h12195-1.pdf

Data Quality Profile for the CE

Internal

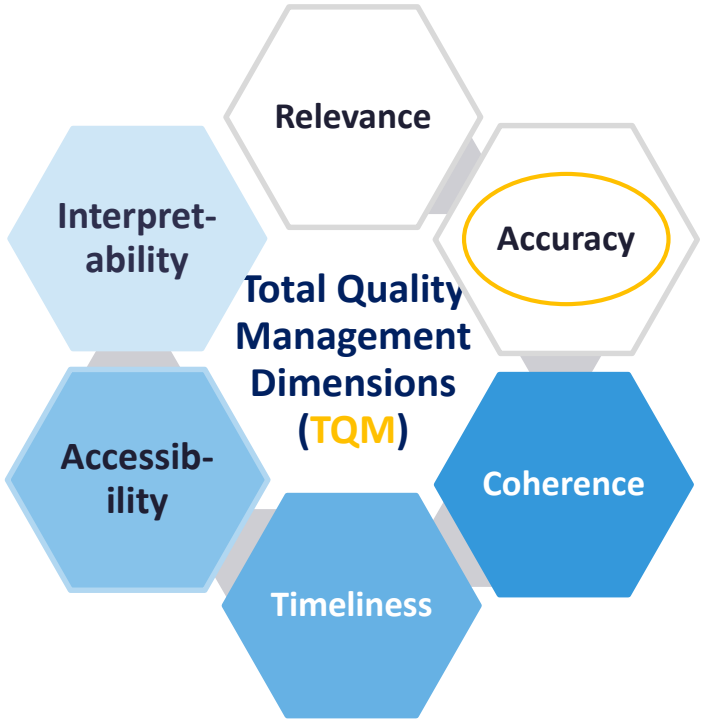
External

“Monitoring; Establish baselines”

“Fitness for Use”

Definition of Data Quality for CE

Multi-dimensional Definition of Data Quality adopted for CE



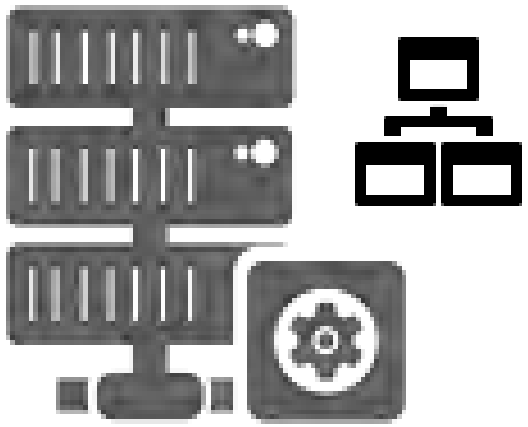
Total Survey Error Sources (TSE)	
Frame (coverage)	*Specification (construct)*
Sampling	*Measurement*
Non-response	*Processing (data edit)*
Post-survey adjustment	

(Gonzalez et al 2009)
https://www.bls.gov/cex/ovrvwdataqualityrpt.pdf

Challenges

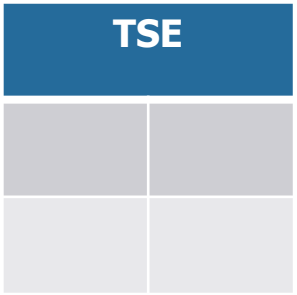
To achieve reproducibility and interpretability *of metrics*

Metric Documentation:
efficient and robust



Infrastructure :
Continuous and adaptable
to change

CE DQP Challenges

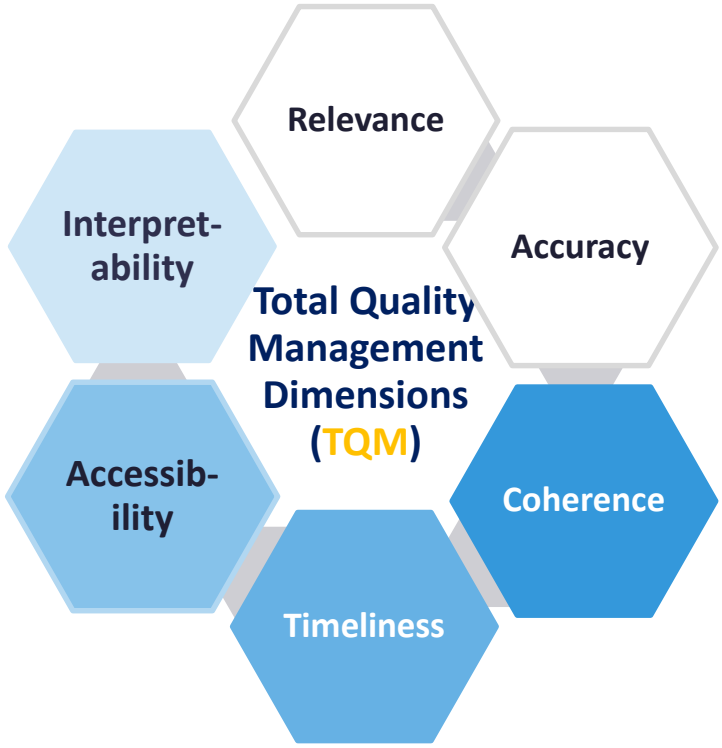


1. Requires participation and coordination across the survey program

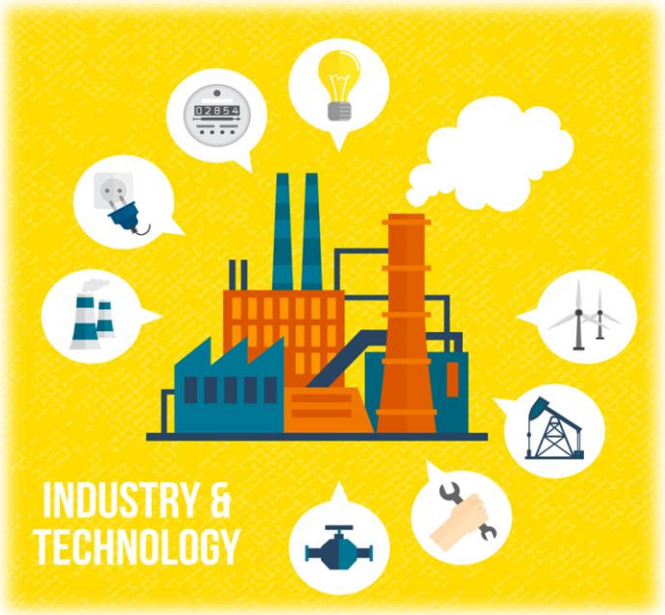
2. Resource intensive to develop and maintain

CE Strategy to identify metrics

TQM: Survey as a manufacturing process



=



http://www.freepik.com/free-vector/industry-and-technology-background_1048768.htm Designed by Freepik

Proposed Framework

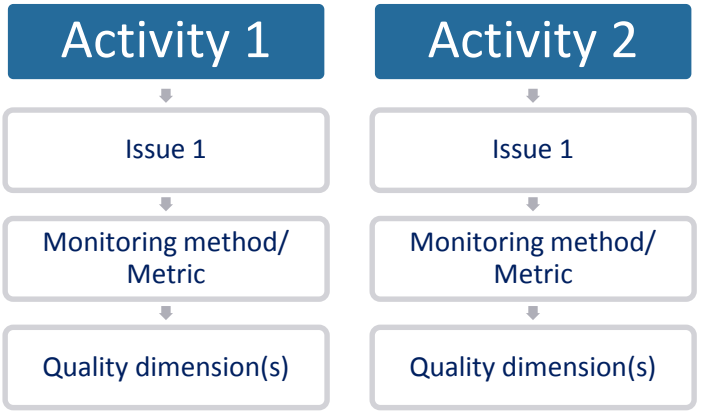
Identifying key stages in CE life cycle

For each stage, identify major activity

For each activity, identify issue(s) of concern

Propose how to monitor issue identified

Identify quality dimension(s) affected



(Fricker et al 2012)

Example of metric metadata description using a template

- Metric Name
- Description
- Metric interpretation
- Survey
- Quality dimension

CALCULATION

- Formula
- Data source and variables
- Frequency
- Level of aggregation
- Maintained by

MONITORING

- Target / Threshold / Tolerance
- Presentation / display

NOTES/COMMENTS

Proposed framework:
Criteria for Metric Prioritization

S.M.A.R.T

Specific – targeted at identified risk

Measurable – can be used to determined progress

Achievable – realistically attainable

Relevant – not just "good to know", actionable

Timely – available when needed

Iterative approach to DQP development

"Learn by doing, Refine and Scale up!"

LESSONS LEARNED

In 2009, DQ definition Adopted for CE

2011

- UNDERSTAND THE TASK FOR WHICH WE WANT TO DEVELOP METRIC
- IMPORTANCE OF METRIC METATDATA DOCUMENTATION FOR REPRODUCIBILITY AND INTERPRETATION OVER TIME

2012

- PROPOSE FRAMEWORK FOR DQP
- ENSURE CONSISTENCY IN DOCUMENTING KEY ELEMENTS OF METRIC METADATA

→ USE OF A TEMPLATE

2013-14

MEASUREMENT ERROR STUDY (WESTAT CONTRACT)

- NO SINGLE **"BEST"** METHOD

→ MULTIPLE METHOD AND INDICATORS (MMI) APPROACH

2015

DQP VERSION 1

- RESPONSE RATES AND EDIT RATES

2016

MMI FOLLOW-UP

- EXTERNAL INDICATORS FEASIBILITY STUDY

2017

DQP VERSION 2 IN PROGRESS

LESSONS LEARNED

In 2009, DQ definition Adopted for CE

2011

- UNDERSTAND THE TASK FOR WHICH WE WANT TO DEVELOP METRIC
- IMPORTANCE OF METRIC METATDATA DOCUMENTATION FOR REPRODUCIBILITY AND INTERPRETATION OVER TIME

2012

- **PROPOSE FRAMEWORK FOR DQP**
- **ENSURE CONSISTENCY IN DOCUMENTING KEY ELEMENTS OF METRIC METADATA**

➡ **USE OF A TEMPLATE**

2013-14

MEASUREMENT ERROR STUDY (WESTAT CONTRACT)

- NO SINGLE **"BEST"** METHOD

➡ MULTIPLE METHOD AND INDICATORS (MMI) APPROACH

2015

DQP VERSION 1

- RESPONSE RATES AND EDIT RATES

2016

MMI FOLLOW-UP

- EXTERNAL INDICATORS FEASIBILITY STUDY

2017

DQP VERSION 2 IN PROGRESS

LESSONS LEARNED

In 2009, DQ definition Adopted for CE

2011

- UNDERSTAND THE TASK FOR WHICH WE WANT TO DEVELOP METRIC
- IMPORTANCE OF METRIC METATDATA DOCUMENTATION FOR REPRODUCIBILITY AND INTERPRETATION OVER TIME

2012

- **PROPOSE FRAMEWORK FOR DQP**
- **ENSURE CONSISTENCY IN DOCUMENTING KEY ELEMENTS OF METRIC METADATA**

➡ **USE OF A TEMPLATE**

2013-14

MEASUREMENT ERROR STUDY (WESTAT CONTRACT)

- NO SINGLE **"BEST"** METHOD

➡ MULTIPLE METHOD AND INDICATORS (MMI) APPROACH

2015

DQP VERSION 1

- **RESPONSE RATES AND EDIT RATES**

2016

MMI FOLLOW-UP

- EXTERNAL INDICATORS FEASIBILITY STUDY

2017

DQP VERSION 2 IN PROGRESS

Example of CE DQP Version 1

CE Data Quality Report (Prototype)
Laura Erhard and Lucilla Tan
Bureau of Labor Statistics

Overview

The Consumer Expenditure Survey (CE) has historically provided some limited metrics for data users to evaluate the overall quality of output provided in its products. Published tables provide standard errors, the public-use microdata user guide provides response rates, and the public-use microdata datasets provide all the variables and flags necessary for users to create his or her own quality measures. There has long been a recognition for the need for more comprehensive data quality metrics that are timely and routinely updated, accessible to data users from a single source. However, there is also recognition of the high cost in terms of resources and

1. **Response Rates**
2. **Nonresponse rates**
3. **Expenditure Edit Rates**
4. **Income Imputation rates**

* Reporting period: 2010 - 2013

https://www.bls.gov/cex/ce_dqreport.pdf

LESSONS LEARNED

In 2009, DQ definition Adopted for CE

2011

- UNDERSTAND THE TASK FOR WHICH WE WANT TO DEVELOP METRIC
- IMPORTANCE OF METRIC METATDATA DOCUMENTATION FOR REPRODUCIBILITY AND INTERPRETATION OVER TIME

2012

- **PROPOSE FRAMEWORK FOR DQP**
- **ENSURE CONSISTENCY IN DOCUMENTING KEY ELEMENTS OF METRIC METADATA**

➡ **USE OF A TEMPLATE**

2013-14

MEASUREMENT ERROR STUDY (WESTAT CONTRACT)

- NO SINGLE **"BEST"** METHOD

➡ MULTIPLE METHOD AND INDICATORS (MMI) APPROACH

2015

DQP VERSION 1

- RESPONSE RATES AND EDIT RATES

2016

MMI FOLLOW-UP

- EXTERNAL INDICATORS FEASIBILITY STUDY

2017

DQP VERSION 2 IN PROGRESS

CE DQP Version 2

Consumer Expenditure Survey

Data Quality Profile Prototype (iteration 2: INTERNAL REPORT)

Evan Hubener, Clayton Knappenberger, Julie Sullivan, & Lucilla Tan **(draft 2017.06.27)**

Overview

The Consumer Expenditure Survey (CE) has historically provided some limited metrics for data users to evaluate the overall quality of output provided in its products. Published tables provide standard errors; the public-use microdata user guide provides response rates, and the datasets contained in the public-use microdata provide all the variables and flags necessary for users to create his or her own quality measures. There has long been a recognition for the need for more comprehensive data quality metrics that are timely, routinely updated, and accessible to data users from a single source, a Data Quality Profile (DQP). However, there is also recognition of the high cost in terms of resources and commitment to identifying appropriate metrics and establishing the information base necessary to routinely

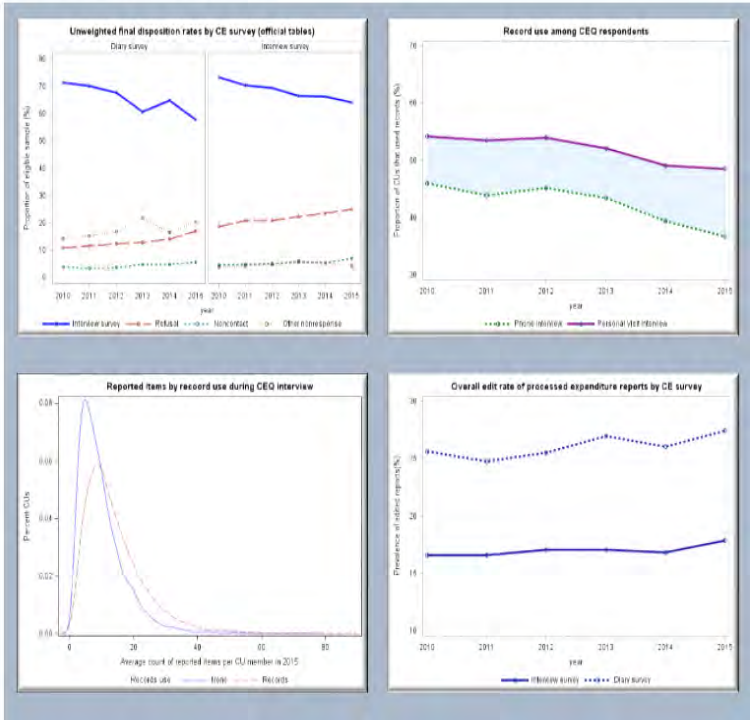
Content links

Visual Summary

Metrics:

- Response rates: official published tables
- Response rates: collected data *
- Use of Records in the CEQ
- Expenditures edit rate: processed data
- Expenditures edit rate: reported data *
- Income Imputation Rates

Figure 1. Select metric trends from 2010 to 2015



DQP Version 2: Scale up from DQP version 1

Contents

- Updated metric reporting period: 2010-2015
- New metric added: Use of Records by Survey Mode
- Metrics refined:
  - Reponses rates: Additional breakouts by collection wave (Internal)
  - Expenditure edit rates: Differentiated between processed and reported data (Internal)
- Addition of visual summary of metric trends

DQP Version 2: Scale up from DQP version 1

- Production Process
  - Coordinated team from 3 areas of the CE Program
  - Use of metric metadata template for Documentation
  - All coding for analysis of metrics and graphs produced within SAS

Moving forward

Lessons Learned from DQP 2

- Spend more time for creating and reviewing the data
- Spend more time for exploring and discussing metric ideas, and document!
- Consult “topic experts”
- Moving the DQP to routine production will need further consideration about the infrastructure needed to support that

Next

- Upcoming: Data Quality Profile version 2 will be available for public users on SEPTEMBER
- We will appreciate your feedbacks and comments!

Contact Information

Yezzi Angi Lee
Economist
Division of Consumer Expenditure Surveys
www.bls.gov/cex
202-691-5154
Lee.Yezzi@bls.gov