NCRM National Centre for Research Methods

Introduction to Survey Data Quality

Olga Maslovskaya

University of Southampton

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

UNIVERSITY OF Southampton

MANCHESTER 1824 The University of Manchester

THE UNIVERSITY OF EDINBURGH

Survey Data

- Vast amounts of survey data are collected for many purposes, including governmental information, public opinion and election surveys, advertising and market research as well as scientific research

- Survey data underlie many public policy and business decisions

- Good quality data reduces the risk of poor policies and decisions and is of crucial importance

NCRM National Centre for Research Methods

Survey challenges

- Budgets are severely constrained (survey costs)
- Pressures on providing timely data are greater in the digital age
- Public interest in participating in surveys is declining and now at all time low (response rates)
- When cooperation obtained from reluctant respondents, responses may be less accurate
- New modes of data collection introduce new concerns for data quality (errors)

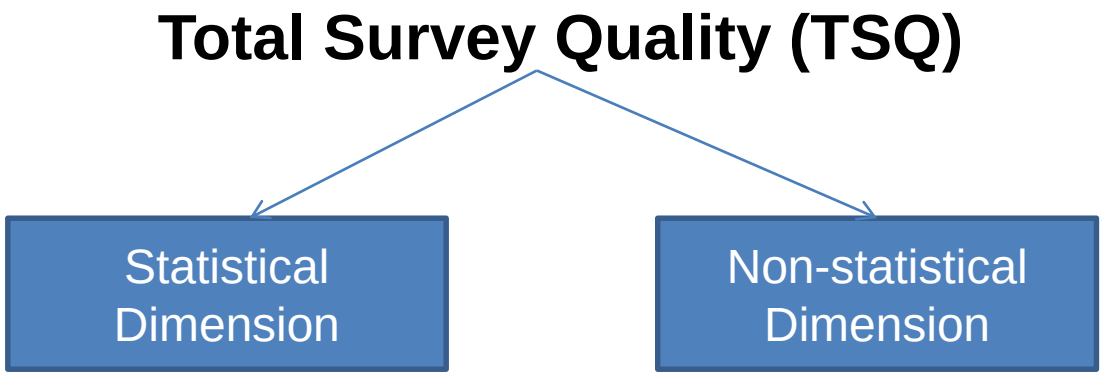
NCRM National Centre for Research Methods

Definition

- **Quality** can be defined simply as “fitness for use”
- **Quality** is a requirement for survey data to be as *accurate* as necessary to achieve their intended purposes, be available at the time it is needed (*timely*), and be *accessible* to those for whom the survey was conducted.

Biemer and Lyberg (2003)

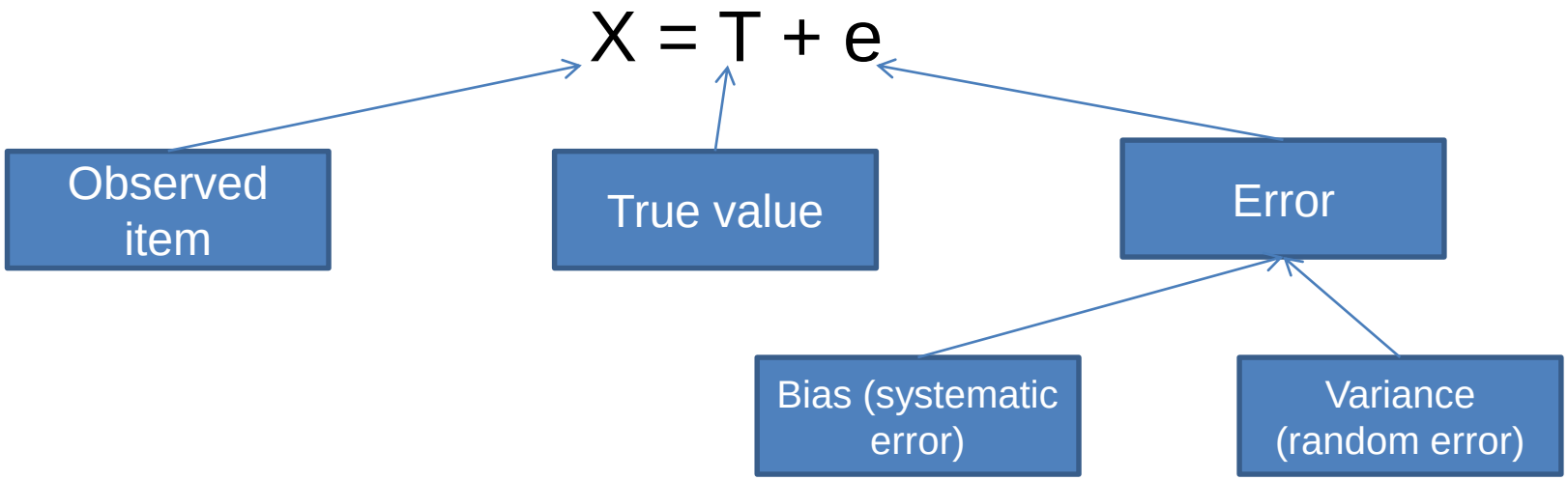
Total Survey Quality (TSQ)



TSQ – survey quality is more than its accuracy or statistical dimension. It also includes among other factors producing results that fit the needs of the survey users and providing results that users will have confidence in. Usability of results is of crucial importance.

TSQ: Quality Dimensions –Statistical

- **Accuracy** of estimates is the difference between the estimate and the true parameter value
- **Accuracy** is the larger concept of TSQ



TSQ: Quality Dimensions – Non-statistical

- **Relevance -** product is relevant and meets user needs
- **Timeliness and punctuality** – in disseminating results (most important user needs)
- **Accessibility and clarity –** of the information
- **Comparability** - reliable comparisons across space and time are often crucial; cross-national comparisons
- **Coherence -** single source – elementary concepts can be combined in more complex ways; different sources – based on common definitions, classifications and methodological standards
- **Completeness -** data are rich enough to satisfy the analysis objectives

TSQ: Quality Dimensions – Non-statistical

- **Credibility** – credible methodology
- **Interpretability –** documentation is clear
- **Richness of detail -** data are rich enough to satisfy the analysis objectives
- **Level of confidentiality protection**
- **Cost –** data give good value for money

Total Survey Error (TSE)

- **TSE** concept was developed by Robert Groves (1989) in book on Survey Errors and Survey Costs
- Survey estimates are derived from complex survey data, published estimates may differ from their true parameter values due to survey errors
- Total Survey Error is the difference between a population mean, total, or other population parameter and the estimate of the parameter based on the sample survey (or census) (Biemer and Lyberg, 2003)

TSE

TSE= sampling errors + non-sampling errors

Sources of Sampling Error

Sampling errors – can be computed for probability samples and are due to selecting a sample instead of the entire population

Sources:

- **Sampling scheme**
- **Sample size**
- **Estimator choice**

Components of Non-sampling Error

Non-sampling errors (including measurement error – cannot be formally estimated but can be improved by interviewing procedures and question wordings etc.) - are errors due to mistakes or system deficiencies, also from incomplete responses to the survey or its questions, etc.

1. Specification error
2. Frame error
3. Nonresponse error
4. Measurement error
5. Data processing error
6. Modelling/Estimation error

Other Important Factors

A number of additional factors can impact survey data quality. Four of the more important:

- the length of time the survey was fielded,
- the use of incentives,
- the reputation of the organisation conducting the survey
- mode of data collection

Actors affecting data quality

- **Respondents** (respondents' effects on data quality): e.g., response styles, satisficing (less efforts to provide optimal responses)
- **Interviewers** (interviewers' effects on data quality): e.g., fabrication, ability to elicit interest and commitment to survey in respondents, duration of interview, duplication of responses apart from say demographic
- **Supervisors and survey research organisations** (supervisors' effects on data quality), e.g. sampling design, training of field workers

NCRM National Centre for Research Methods

Data quality monitoring strategies

- **Continuous quality improvement (CQI)**
  - Special cause variation – errors made by individual coders
  - Common cause variation – errors due to the process itself
- **Responsive design** (Groves and Heeringa, 2006)
- **Adaptive design** – real time control of costs and errors – (Schouten et al. 2013)
- **Adaptive Total Design** (Biemer, 2010) – adaptive design strategy that combines ides of CQI and the TSE paradigm to reduce costs and error across multiple survey processes
- **Six Sigma** – set of principles and strategies for improving any process

Data Quality in Practice

- No instance where a total survey quality (TSQ) measure has ever been calculated or combined single measure of quality taking all dimensions into account
- Cost-benefit trade-offs to **minimise** different errors depending on survey aims
- Q**uality reports** or **quality declarations** have been used where information on each dimension is provided
- **Data quality guides** are meant to alert the data user to potential sources of bias that might be present

Conclusions (1)

- Data quality is a multi-dimensional concept
- Single score or measure of data quality is not available
- Cost-benefit trade-offs to minimize different errors depending on survey aims
- Quality frameworks are developed and adopted
- Broad range of relevant data quality indicators and information are available with data
- The chances of users misusing the data or misinterpreting published statistics is reduced if they understand better the strengths and limitations of the data.
- New technologies require fresh considerations of data quality issues in new types of surveys

Conclusions (2)

High quality of the survey data brings

- improvement in the quality of surveys themselves
- improvement of the quality of research and of policy and financial decisions that are based on the survey data

References

- **Biemer** (2010) Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5): 817-848.
- **Biemer** (2016) Total Survey Error Paradigm: Theory and Practice. In *The Sage handbook of survey methodology* by Wolf, Joye, Smith and Fu. London: SAGE publications.
- **Biemer and Lyberg** (2003) *Introduction to survey quality*. New York: John Wiley & Sons.
- **Groves and Heeringa** (2006) Responsive design for household surveys: Tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society Series A*, 169 (3): 439-457.
- **Lynn** (2004) Editorial: Measuring and communicating survey quality. *Journal of the Royal Statistical Society Series A*, 167 (4): 575-578.
- **Lyberg and Weisberg** (2016) *The SAGE handbook of survey methodology*. London: SAGE publications.
- **Schouten et al**. (2013) Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39 (1): 29-39.
- **Weisberg** (2005) *The total survey error ap*proach. Chicago: University of Chicago Press.