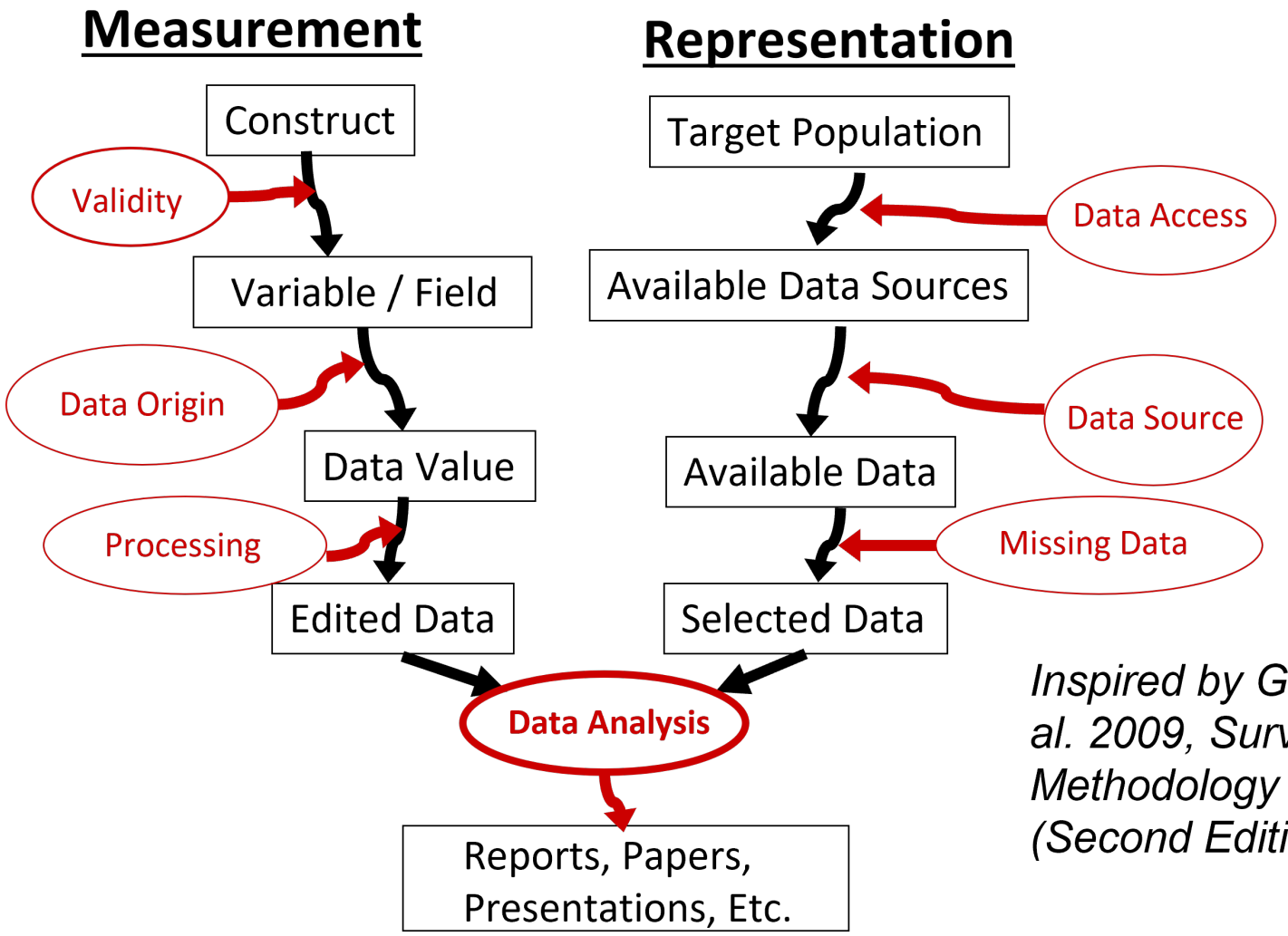
Why is Data Analysis Part of Total Data Quality?

Brady T. West

Dimensions of TDQ: The Big Picture!



Inspired by Groves et al. 2009, Survey Methodology (Second Edition)

High-Quality Data Analysis

- We now focus on the **quality** of the data analysis!
- Data analysis appears **last** in the overall framework and is usually a last step before research results are released to the public / stakeholders.
- A failure to analyze the data using appropriate methods can **negate all prior work done to maximize quality** across the different dimensions of the TDQ framework!

Analyzing Designed Data (1)

- While data collectors often analyze their own data, a very common research tool is secondary analysis of designed data that were collected by other researchers!

- Designed data (and especially survey data) often have analytic plans developed at the time of planning data collection; very important to follow these plans to ensure unbiased inferences about populations of interest

Analyzing Designed Data (2)

- **For example:** survey data often include weights (enabling one to map a sample back to a population, especially if there was oversampling of certain subgroups), and other codes describing features of the sample design.
- If one is doing secondary analyses of the cleaned and processed survey data set, have those analyses accounted for relevant design features (e.g., population weights and sample design codes)? **We will see examples…**
- Have appropriate statistical models been specified for the types of variables being analyzed? **We will see examples…**

Analyzing Gathered Data

- Statistical models are often fit to **gathered data** to understand tendencies, distributions, relationships, and patterns
- Do these models account for appropriate covariates / confounders when examining relationships (very important for observational studies, where groups being compared may differ along other dimensions)?
  - Also, is **over-fitting** a problem?
- Have appropriate models been specified for the types of variables being analyzed (valid for both types of data)?
- Are there biases inherent to the algorithms being used when applying machine learning techniques?

What's Next?

- We will discuss important threats to the quality of the analysis of designed data.
- We will read an article talking about the prevalence of analytic error in secondary analyses of survey data, and the implications of such errors.
- Then, we will turn to analytic considerations for gathered data.

UNIVERSITY OF MICHIGAN