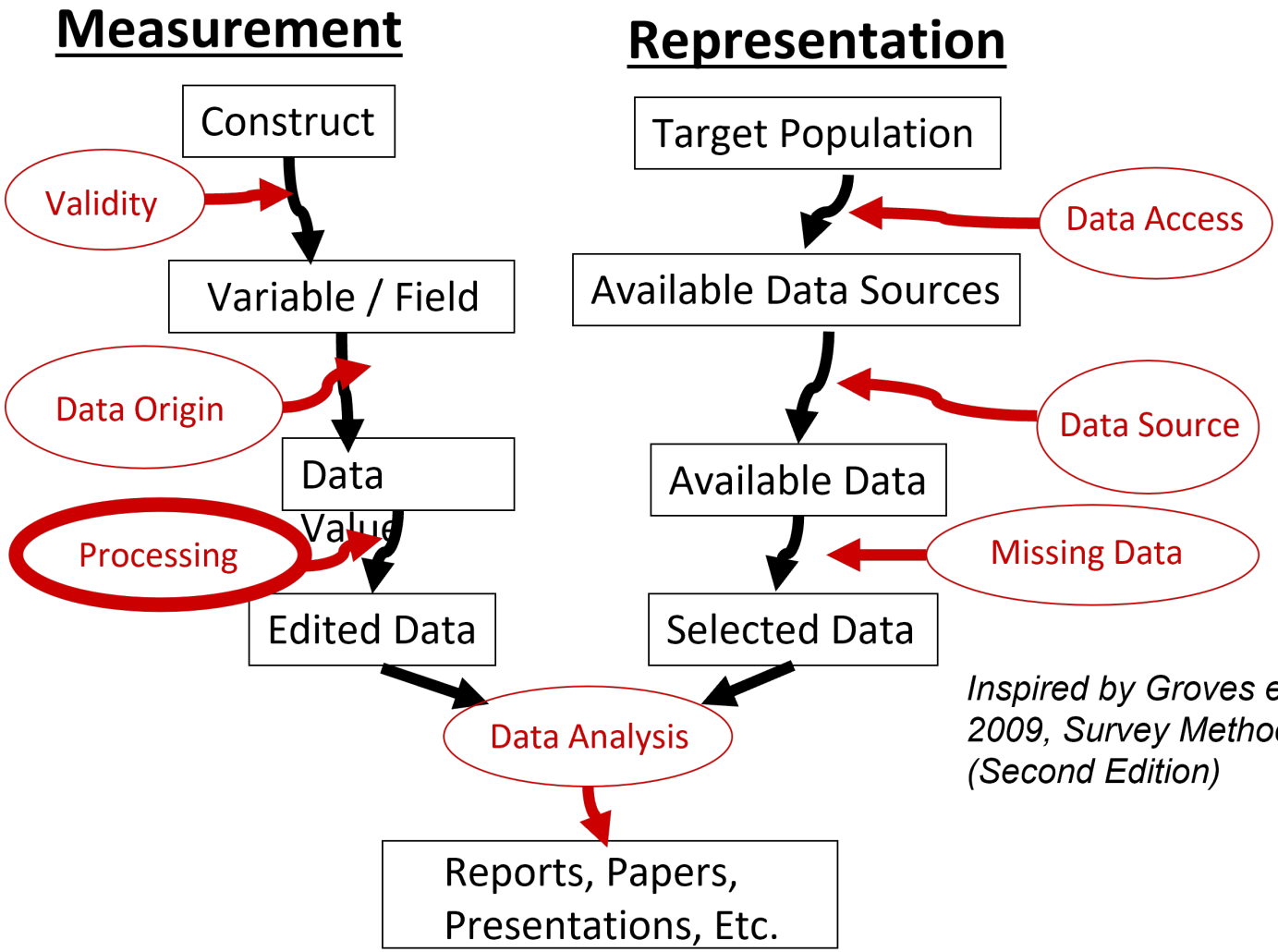
Data Processing Threats for Designed Data

By Brady T. West

Dimensions of TDQ: The Big Picture!



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

Data Processing Threats for Designed Data (1)

- Incorrect edit rules / recodes applied when cleaning and preparing survey data sets for analysis (e.g., recoding missing data into a valid category)
- Poor data linkage procedures (e.g., mismatches when linking survey data with administrative data), incompatibility in linkage keys, missing data on linkage keys
  - Probabilistic linkage techniques can reduce error but are not perfect!
- Ill-specified models for imputation of missing values on variables (e.g., were model diagnostics checked?)

Data Processing Threats for Designed Data (2)

- Incorrect calculation of survey weights / failure to smooth survey weights and reduce impact of outliers
- **Sana and Weinreb (2008)** reported that interviewers did a better job of editing data than data managers!
  - Why? They are closer to the data and make better decisions regarding critical edits.
- **Granquist and Kovar (1997)** found that while a large portion of a federal statistical agency's budget goes to data editing, this generally leads to small changes in estimates.

Data Processing Threats for Designed Data (3)

- **Campanelli et al. (1997):** While between-coder variance in the coding of open-ended occupations tended to be small, its effects on the variance of survey estimates can be inflated with large workloads!
- Also, computer-assisted coding methods did not offer large reliability improvements over manual coding by professional coders.

What's Next?

- We'll look at a case study of coder variance in a real-world survey.
- We will then turn to a discussion of data processing threats for **gathered data.**

UNIVERSITY OF MICHIGAN