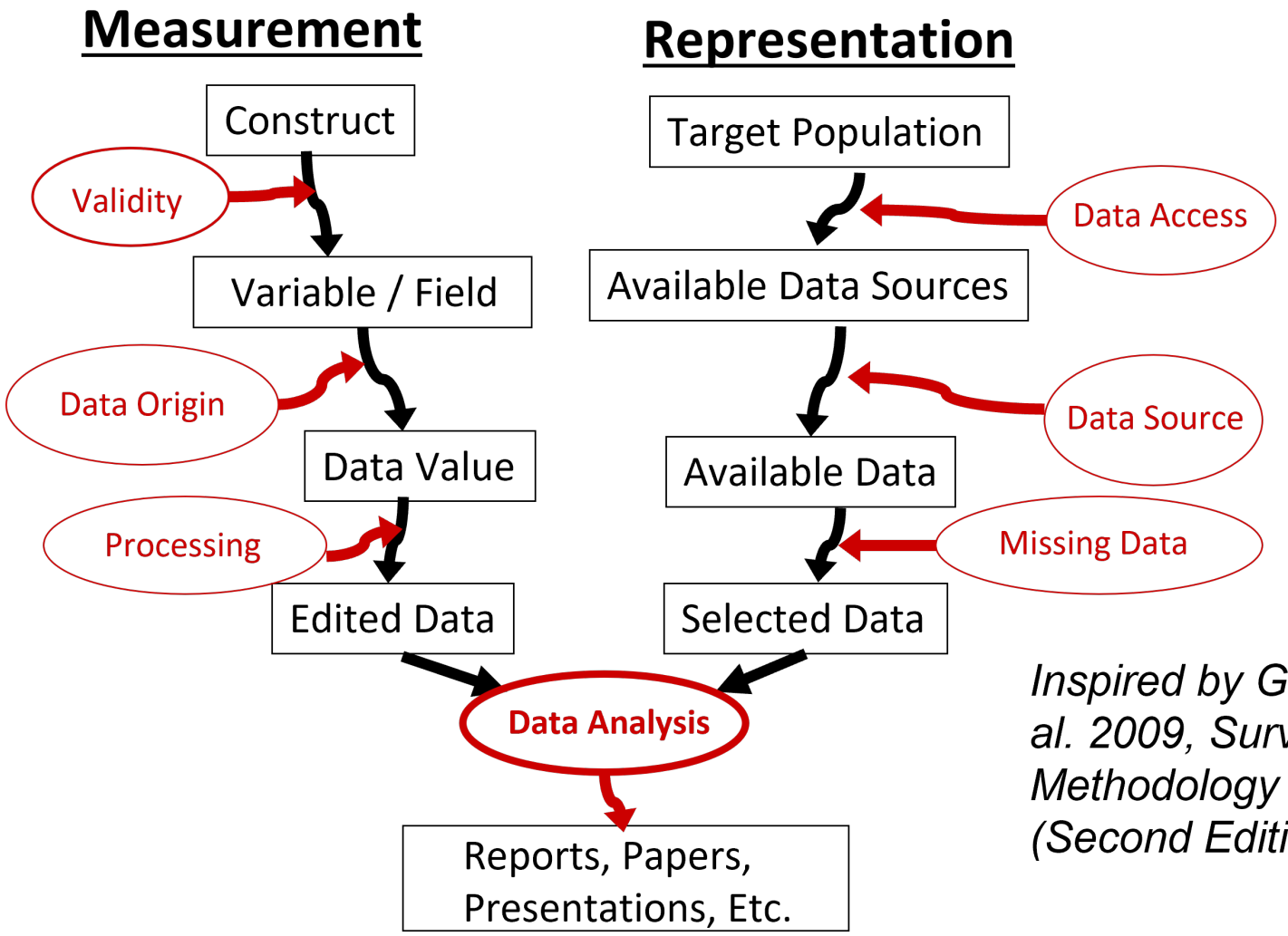
Threats to the Quality of Data Analysis for Designed Data

By Brady T. West

Dimensions of TDQ: The Big Picture!



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

Threats Concerning Data Analysis for Designed Data (1)

- A failure to use **survey weights** and **sampling error codes** in data analysis, which
  - correct for selection bias (data missingness!) due to sampling and nonresponse (survey weights)
  - make sure that standard errors of estimates reflect the expected sampling variability under a given sample design (survey weights and sampling error codes)

  can severely affect inferences based on survey data!
- See the **West et al. (2016)** reading for this week…

Threats Concerning Data Analysis for Designed Data (2)

- Poor specification of statistical models for the data (**Pfeffermann 2011**; **Heeringa et al. 2017**).
- Failure to use appropriate survey weights when fitting **multilevel models** to survey data (**Pfeffermann et al. 1998**).
  - Need survey weights for **both** the units of analysis at Level 1 (typically survey respondents) and the larger clusters at Level 2 and above (e.g., sampled neighborhoods)

Examples: Data Analysis Threats for Designed Data (1)

- Apparent analytic errors in secondary analyses of survey data are seemingly widespread (**West et al. 2016**)
- An extreme example from the 2010 National Survey of College Graduates (**West et al. 2016**):
  - Estimated % with Primary Job in Science and Engineering (key indicator!):
    - **Fully accounting for final survey weights, strata, cluster sampling:**
      - 30.38% (SE = 0.30%)
    - **Accounting for final survey weights only:**
      - 30.38% (SE = 0.39%; *no gains in SE from stratification!*)
    - **Completely ignoring sample design features:**
      - 54.94% (SE = 0.20%) (!!!!!)

Examples: Data Analysis Threats for Designed Data (2)

- **West and Sakshaug (2018)** find that analytic errors also seem to be quite prevalent in analyses of establishment survey data, with similar implications for inference.
- Analytic errors also seem to be common in the medical field (**Khera et al. 2017**).
- **Korn and Graubard (1999)** illustrate the importance of correct model specification and careful consideration of design features using real data.

What's Next?

- We will read an article talking about the prevalence of analytic error in secondary analyses of survey data, and the implications of such errors (**West et al., 2016**).
- Next, you will have access to an optional (but highly recommended!) tutorial on how to use the free R software.
  - Feel free to go through as few or as many modules as desired!
- We will then see a live demonstration, using R, of alternative approaches to the analysis of designed survey data in practice.
- Then, we will turn to analytic considerations for gathered data.

UNIVERSITY OF MICHIGAN