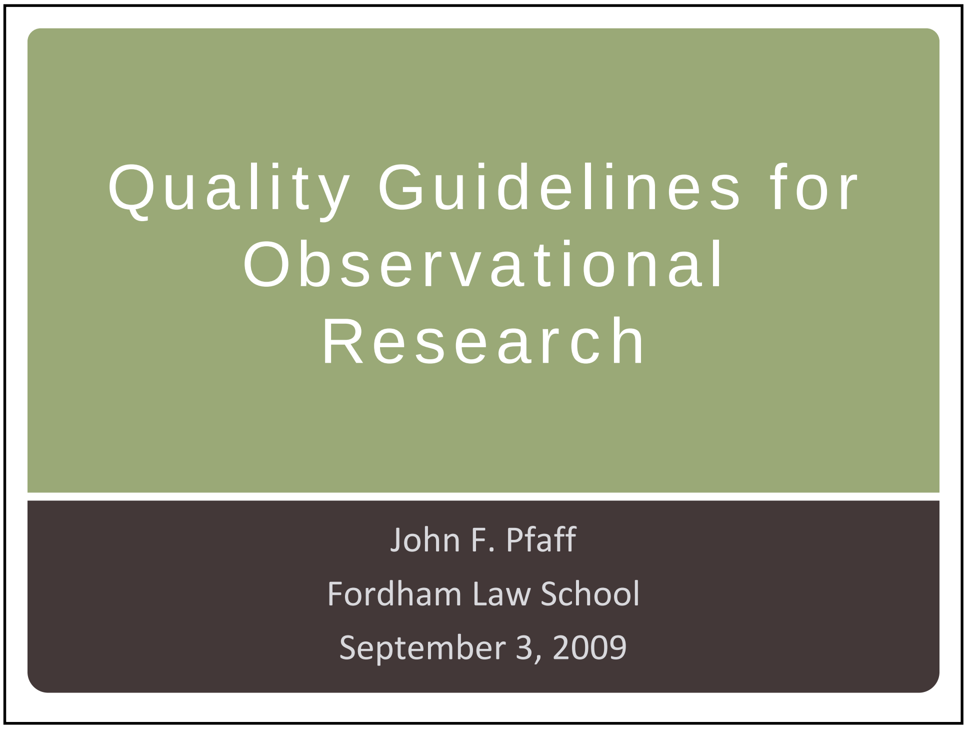
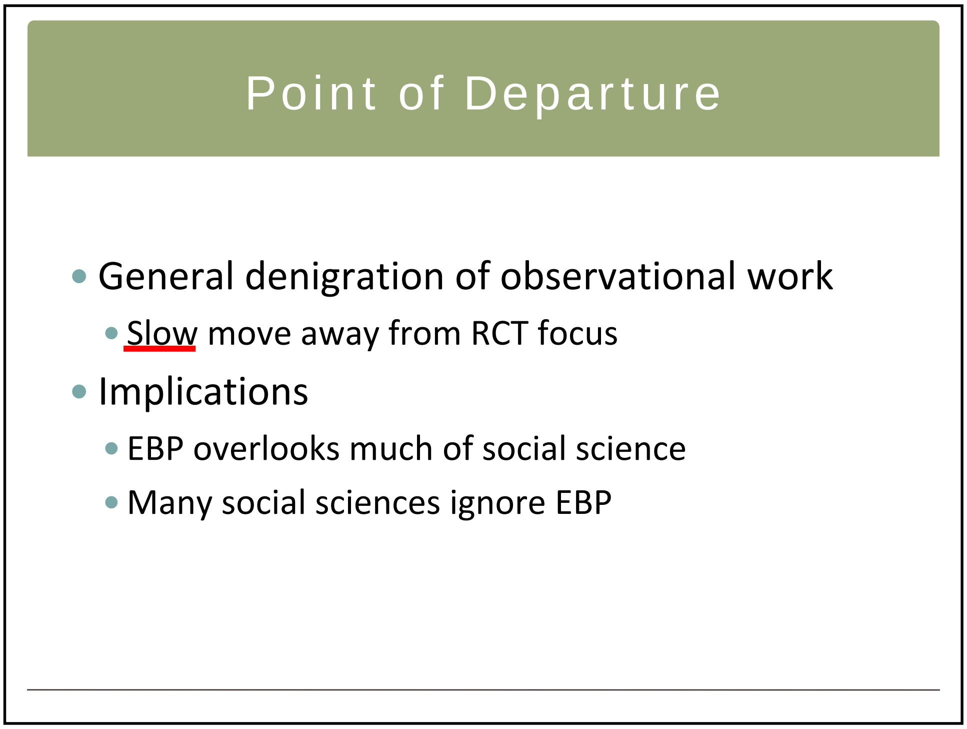


Quality Guidelines for Observational Research

John F. Pfaff
Fordham Law School

September 3, 2009



Point of Departure

- General denigration of observational work
  - Slow move away from RCT focus
- Implications
  - EBP overlooks much of social science
  - Many social sciences ignore EBP

Goals

- **Why we need guidelines for observational work**
  - RCTs are not always feasible
  - Observational work much riskier
- **What these guidelines will look like**
  - Can't just import: RCT model inapplicable
  - Key challenges:
    - Difficult normative questions about defining quality
    - Deeper explorations of threats to quality
    - Substantially more complex to design

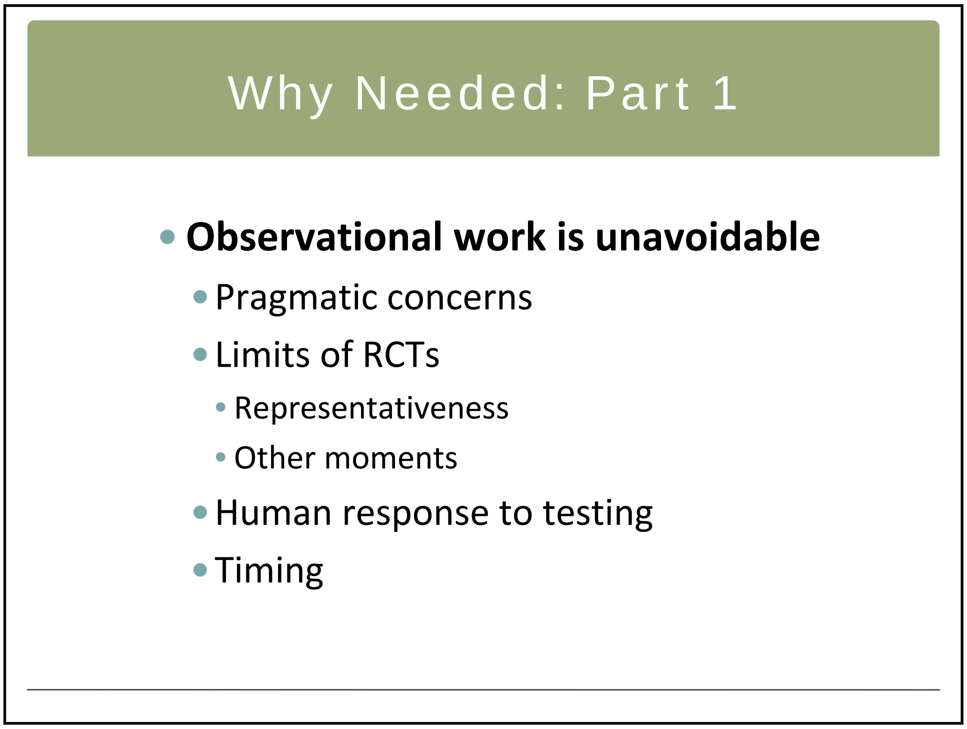
The Gauntlet

social scientists

"If ~~epidemiologists~~ cannot define what constitutes quality in non-experimental studies, how is it possible to do studies that we all agree have merit? If meta-analysis fails because quality is elusive, then all of non-experimental ~~epidemiology~~ fails for the same reason."

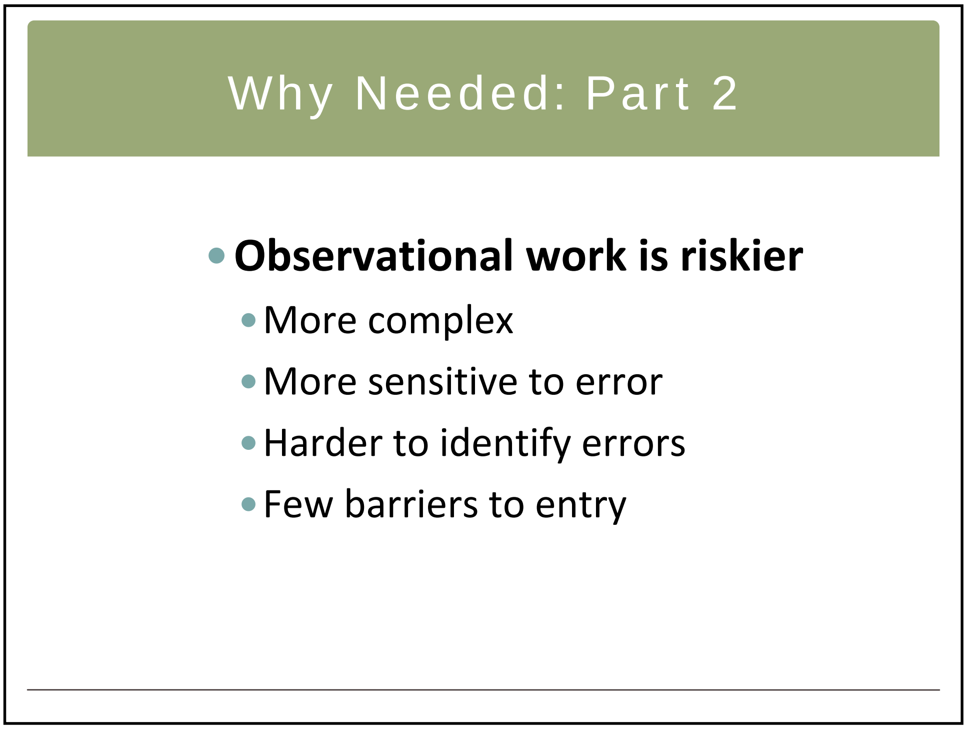
social science

 Diana Petitti, 1994. "Of Babies and Bathwater," *Am J Epidemiology* 140: 779-782.



Why Needed: Part 1

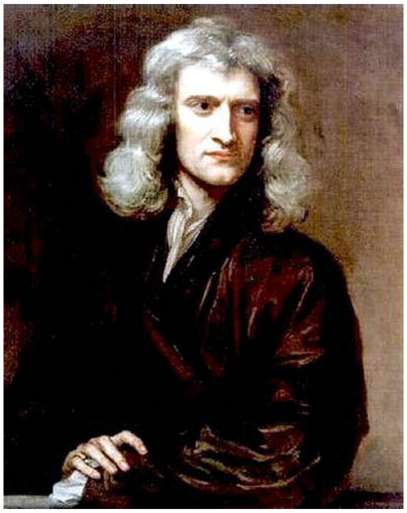
- **Observational work is unavoidable**
  - Pragmatic concerns
  - Limits of RCTs
    - Representativeness
    - Other moments
  - Human response to testing
  - Timing



Why Needed: Part 2

- **Observational work is riskier**
  - More complex
  - More sensitive to error
  - Harder to identify errors
  - Few barriers to entry

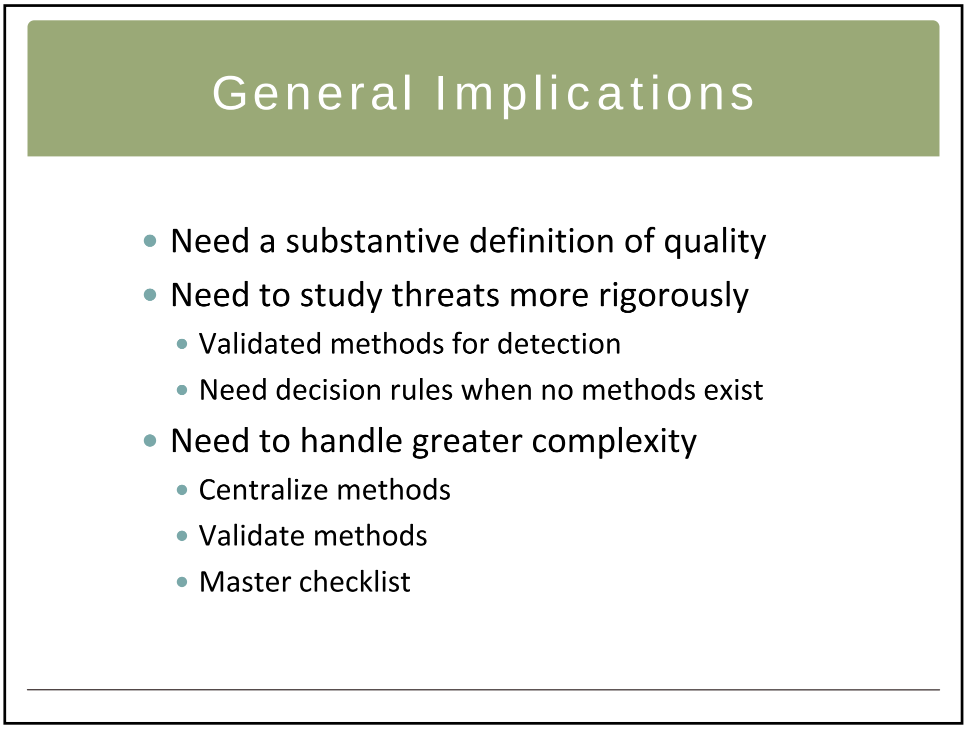
Newton's Third Law



For every observational finding, there is an opposite—though not necessarily equal—finding.

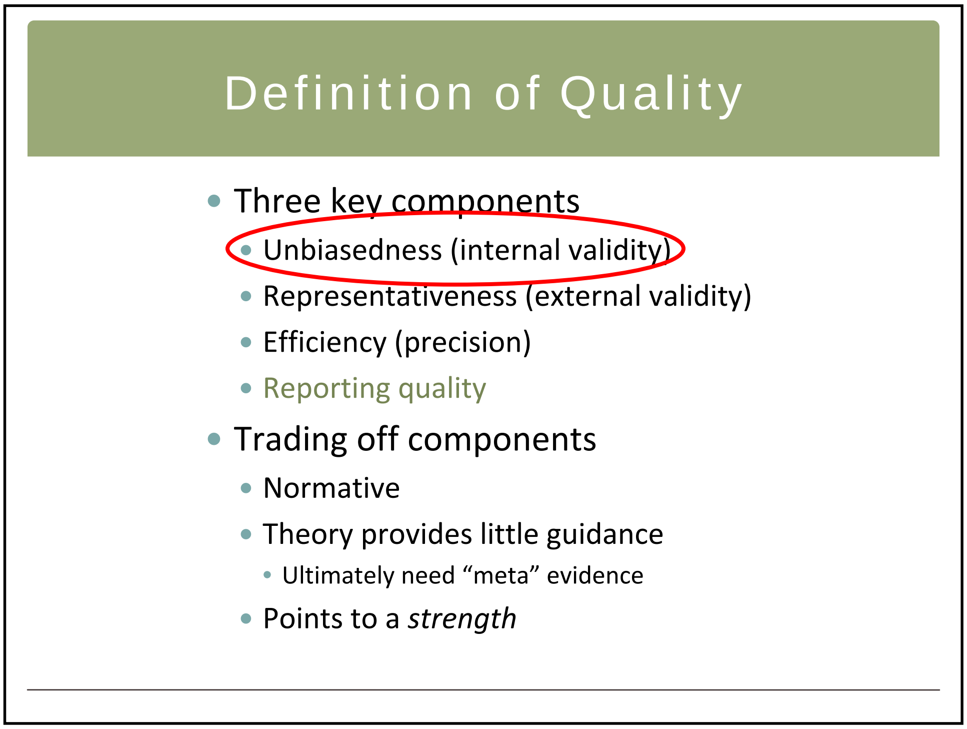
Core Problem

- Obvious need for observational guidelines
- Key: Cannot transfer RCT guidelines
  - Methodological differences
    - RCTs and OR use different methods
    - OR often has competing treatments per threat
  - Procedural vs. substantive difference
    - RCT: One method targets most threats
    - Obs: Each threat has own treatment
    - Obs: Treatment for one threat can aggravate another



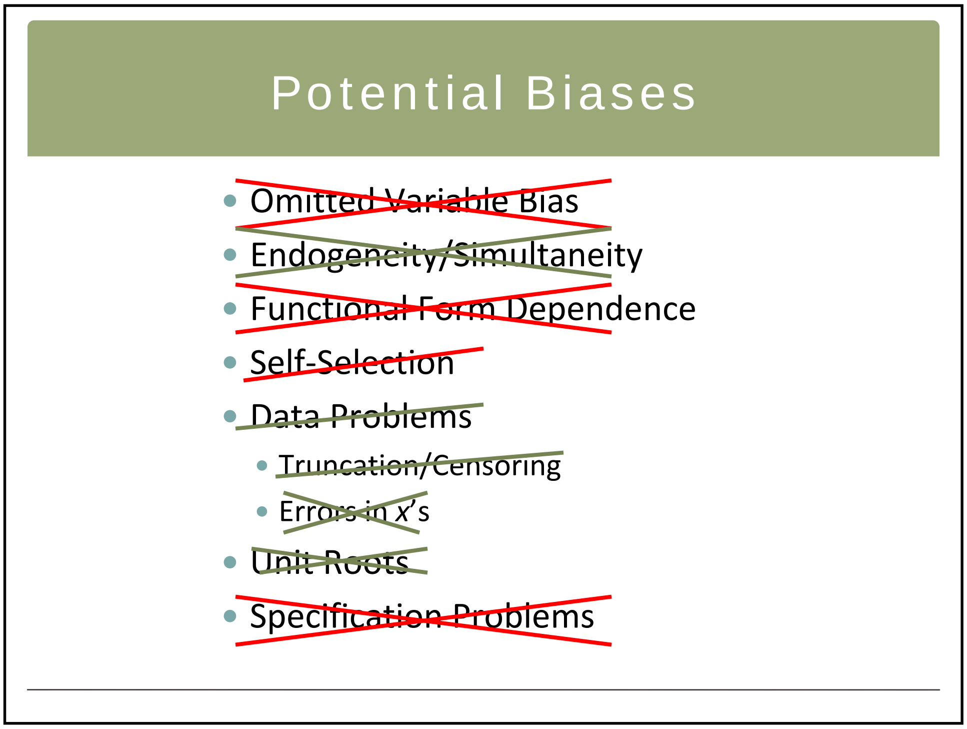
General Implications

- Need a substantive definition of quality
- Need to study threats more rigorously
  - Validated methods for detection
  - Need decision rules when no methods exist
- Need to handle greater complexity
  - Centralize methods
  - Validate methods
  - Master checklist



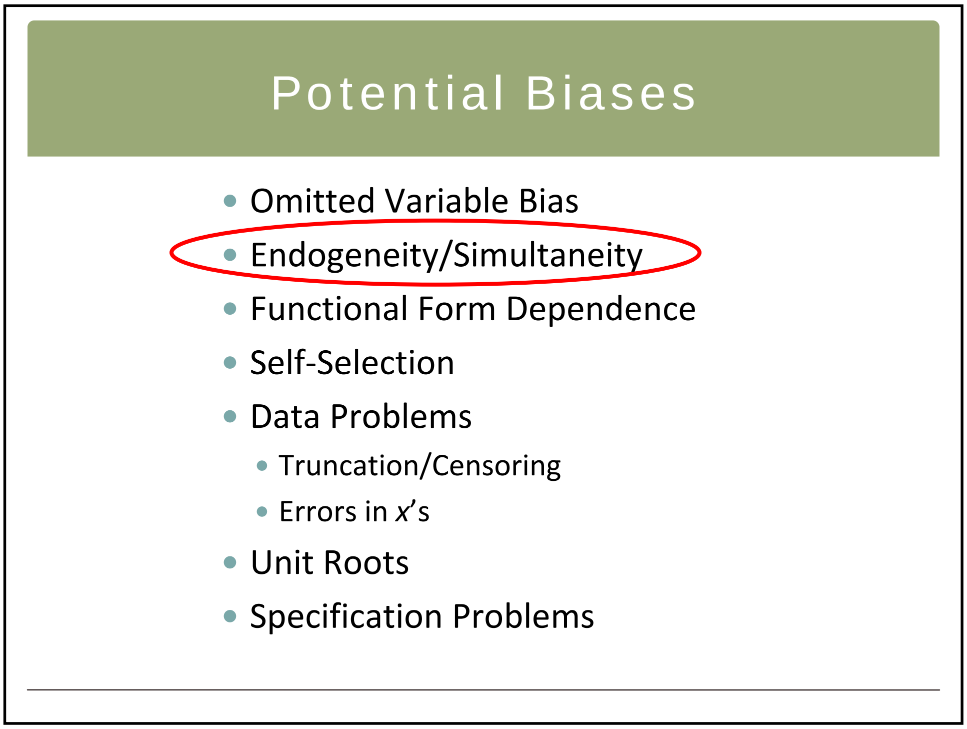
Definition of Quality

- Three key components
  - Unbiasedness (internal validity)
  - Representativeness (external validity)
  - Efficiency (precision)
  - Reporting quality
- Trading off components
  - Normative
  - Theory provides little guidance
    - Ultimately need “meta” evidence
  - Points to a *strength*



Potential Biases

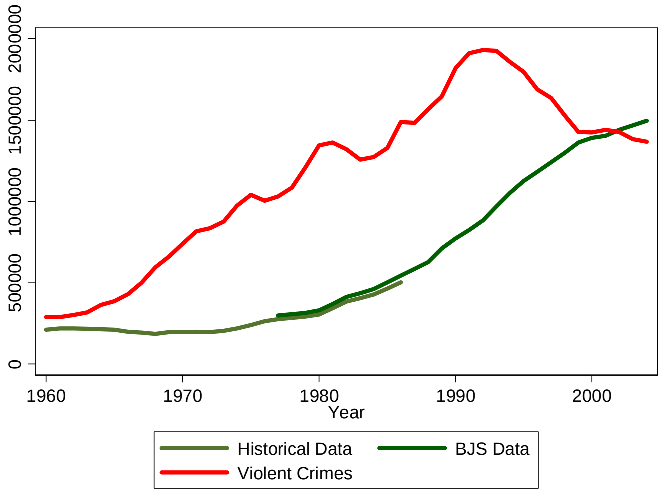
- ~~Omitted Variable Bias~~
- ~~Endogeneity/Simultaneity~~
- ~~Functional Form Dependence~~
- ~~Self-Selection~~
- ~~Data Problems~~
  - ~~Truncation/Censoring~~
  - ~~Errors in *x*’s~~
- ~~Unit Roots~~
- ~~Specification Problems~~



Potential Biases

- Omitted Variable Bias
- Endogeneity/Simultaneity
- Functional Form Dependence
- Self-Selection
- Data Problems
  - Truncation/Censoring
  - Errors in *x*’s
- Unit Roots
- Specification Problems

Applied Example: Incarceration and Crime



Endogeneity

0. Is endogeneity a problem?

- ~~A. Statistical test~~
  - ~~Granger Causality Test~~
- B. Literature review
- C. Intuition or theory

Endogeneity

0. Is endogeneity a problem?

1. Does the paper properly control for endogeneity?

   A1. Does it use quasi-experimental techniques?

   A2. Does it use a regression discontinuity?

   B. Does it use instrumental variables?

   C. Does it use a system of equations?

   D. Does it use something else?

Endogeneity

0. Is endogeneity a problem?

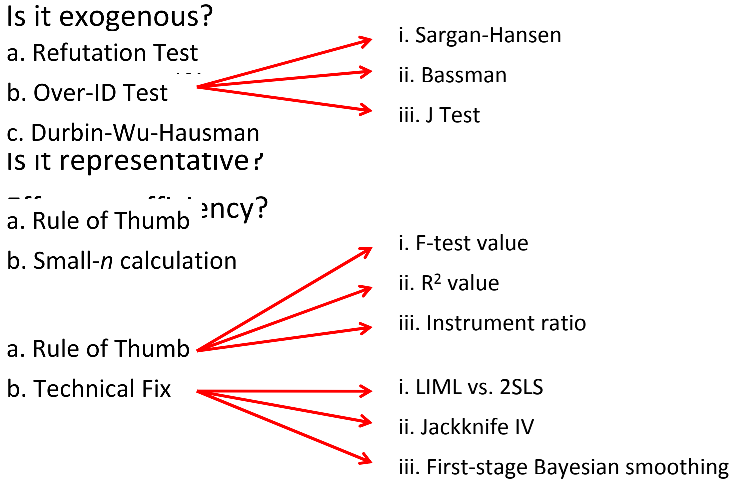
1. Does the paper properly control for endogeneity?

   B. Does it use an instrumental variable?

   1. Is it exogenous?
   2. Is it consistent?
   3. Is it strong?
   4. Is it representative?
   5. Effect on efficiency?

Endogeneity

1. Is it exogenous?
2.
3.
4. Is it representative?
5. ency?



a. Refutation Test

b. Over-ID Test

c. Durbin-Wu-Hausman

i. Sargan-Hansen

ii. Bassman

iii. J Test

a. Rule of Thumb

b. Small-*n* calculation

i. F-test value

ii. R^2 value

iii. Instrument ratio

a. Rule of Thumb

b. Technical Fix

i. LIML vs. 2SLS

ii. Jackknife IV

iii. First-stage Bayesian smoothing

General Implications (Redux)

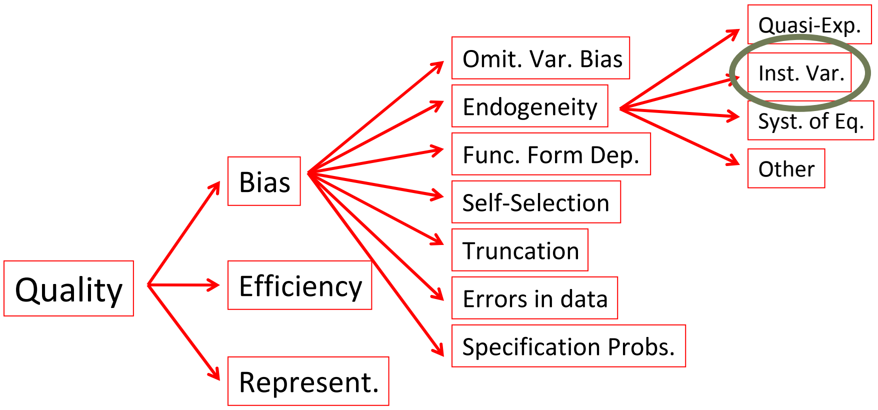
- Need a substantive definition of quality
  - Instruments: bias vs. efficiency and representativeness
  - Normative and subjective
  - Need to develop evidence
    - Theory tells us little
    - Need to develop “meta-evidence”
      - “In general, 2SLS doubles the standard errors”

General Implications (Redux)

- Need rigorous treatment of threats
  - Evidence of threats is“meta”
    - Whether particular method used correctly is internal to paper
    - Whether threat exists is external to paper

General Implications (Redux)

- Handing greater complexity



Quality

Bias

Efficiency

Represent.

Omit. Var. Bias

Endogeneity

Func. Form Dep.

Self-Selection

Truncation

Errors in data

Specification Probs.

Quasi-Exp.

Inst. Var.

Syst. of Eq.

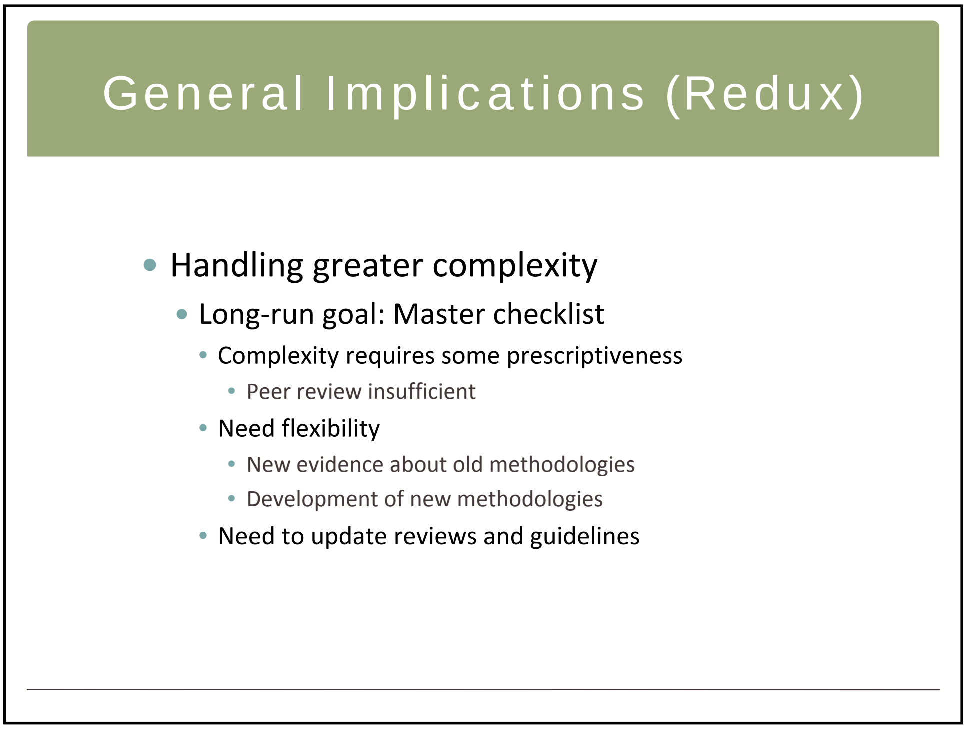
Other

General Implications (Redux)

- Handling greater complexity
  - Short-run goal: centralize methodologies
    - Solutions too widely scattered
      - Theory/practice divide
      - General solutions located in specific substantive articles
    - Need for interdisciplinary collaboration
    - Points to power of internet

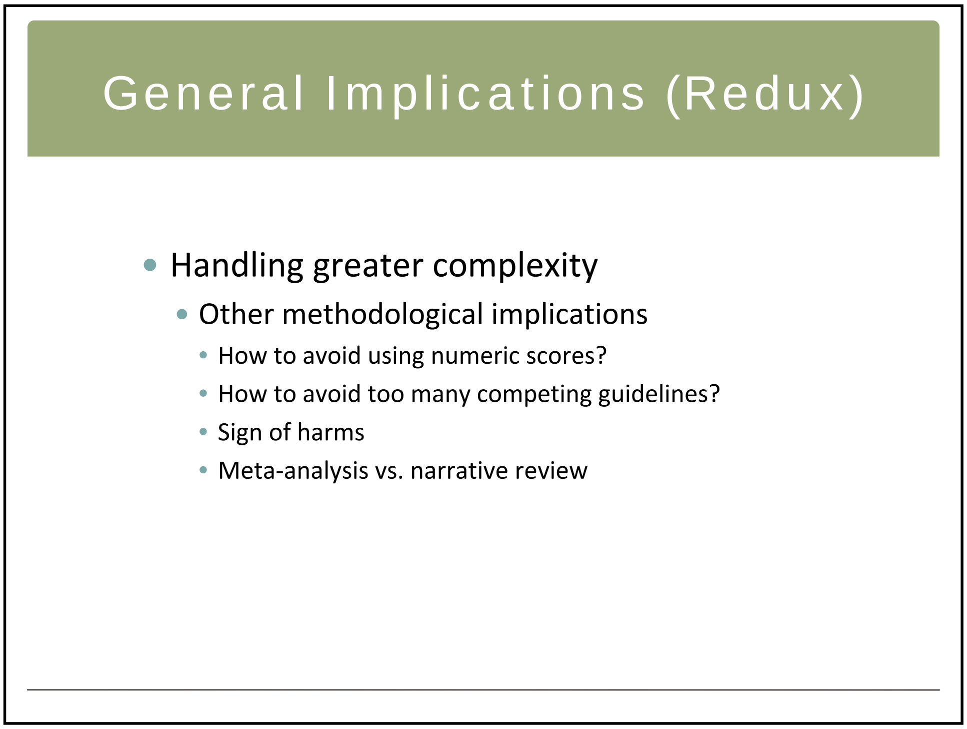
General Implications (Redux)

- Handling greater complexity
  - Intermediate goal: validation
    - Requires within-question literature reviews
      - “To what extent does 2SLS differ from LIML for question *x*?”
    - Then requires review across reviews
      - “Under what conditions does 2SLS differ from LIML?”
    - Time-intensive but necessary
  - Intermediate goal: level of detail



General Implications (Redux)

- Handling greater complexity
  - Long-run goal: Master checklist
    - Complexity requires some prescriptiveness
      - Peer review insufficient
    - Need flexibility
      - New evidence about old methodologies
      - Development of new methodologies
    - Need to update reviews and guidelines



General Implications (Redux)

- Handling greater complexity
  - Other methodological implications
    - How to avoid using numeric scores?
    - How to avoid too many competing guidelines?
    - Sign of harms
    - Meta-analysis vs. narrative review

The Gauntlet

“If ~~epidemiologists~~ social scientists cannot define what constitutes quality in non-experimental studies, how is it possible to do studies that we all agree have merit? If meta-analysis fails because quality is elusive, then all of non-experimental ~~epidemiology~~ social science fails for the same reason.”

— Diana Petitti, 1994. “Of Babies and Bathwater,” *Am J Epidemiology* 140: 779-782.

Guideline Example

1. Does the study control for endogeneity?	Yes Yes
B. Does it use an instrument?	Debatable Yes
1. Is it exogenous?	Yes
a. Refutation test	Yes
b. Over-ID test	Unreported
c. Durbin-Wu-Hausman test	Likely
2. Is it consistent?	51 x 20 Unreported
a. Rule of thumb	•
b. Small-*n* calculation	Debatable
3. Is it strong?	R2 = 0.20 F-stat unreported
a. Rule of thumb	No
b. Technical fix	Debatable
4. Representativeness	SE up 6-fold Debatable