Risk of bias assessment for experimental and quasi-experimental designs based on statistical methods

Hugh Waddington

&

Jorge Garcia Hombrados

"The haphazard way we individually and collectively study the fragility of inferences leaves most of us unconvinced that any inference is believable... It is important we study fragility in a much more systematic way"

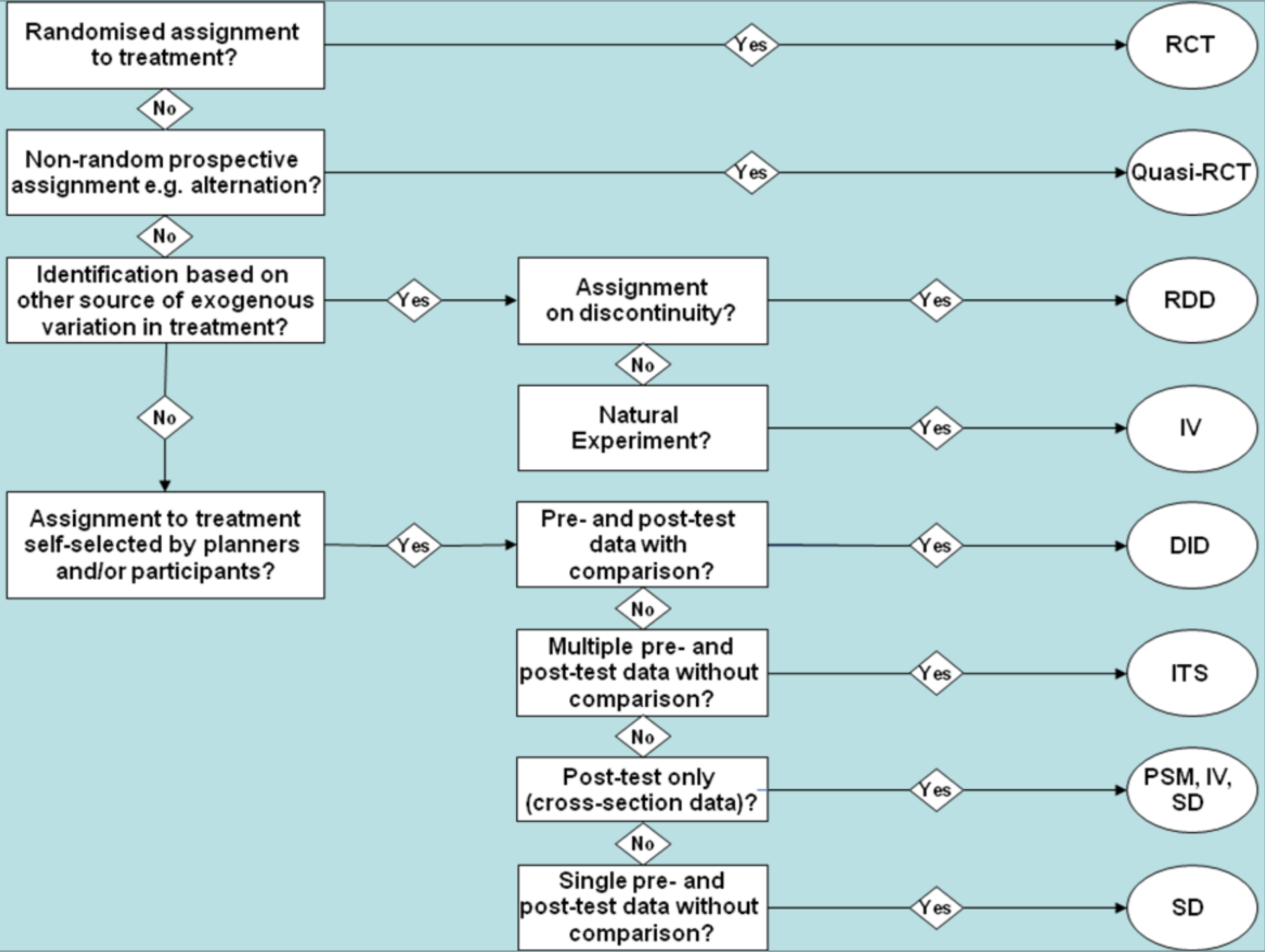
Edward Leamer, 1983: *Let's take the con out of econometrics*

Important elements of study quality

- Internal validity
  - Asks about the causal properties
- External validity
  - Asks about the generalizability of causal properties (or, more generally, the study's findings)
- Construct validity
  - Asks about the language of the measures
- Statistical conclusion validity
  - Asks about the assumptions, estimations, and calculations of summary statistics

E.g. Systematic review by Gaarder et al. “CCTs and health: unpacking the causal chain” Journal of Development Effectiveness, 2010

Programme	Method (authors)
Brazil (Bolsa Familia)	PSM (Morris et al 2004)
Colombia (Familias en Accion)	PSM, DID (Attanasio et al 2005)
Honduras (PRAF)	RCT (Flores et al 2004; Morris et al 2004)
Jamaica (PATH)	RDD (Levy and Ohls 2007)
Malawi (MDICP)	RCT (Thornton 2008)
Mexico (Progresa, Oportunidades)	RCT (Gertler, 2000 etc etc etc), also PSM, RDD, DID
Nepal (SDIP)	ITS (Powell-Jackson et al 2009)
Nicaragua (RPS)	RCT (Barham, Maluccio, Flores 2008)
Paraguay (Tekopora)	PSM (Soares et al 2008)
Turkey (CCT)	RDD (Ahmed et al 2006)


Randomised assignment to treatment?
Yes
RCT
No
Non-random prospective assignment e.g. alternation?
Yes
Quasi-RCT
No
Identification based on other source of exogenous variation in treatment?
Yes
Assignment on discontinuity?
Yes
RDD
No
Natural Experiment?
Yes
IV
No
Assignment to treatment self-selected by planners and/or participants?
Yes
Pre- and post-test data with comparison?
Yes
DID
No
Multiple pre- and post-test data without comparison?
Yes
ITS
No
Post-test only (cross-section data)?
Yes
PSM, IV, SD
No
Single pre- and post-test data without comparison?
Yes
SD

Existing risk of bias tools

- 100s of risk of bias tools exist (Deeks et al. 2003)
- Mainly designed for assessing validity of RCTs and 'epidemiological' designs (controlled before and after designs, cohort designs, case-control designs)
- But do not enable assessment of QEDs commonly used in social enquiry, such as RDDs and IV estimation

Prominent tools

Risk of bias tool	RCTs	Quasi-RCTs	Natural experiment (RDD)	Instrumental variables	Matching (eg PSM)	Diff-in-diff (panel data)	CBA	Cohort	Case-control
AHRQ	√	?						√	√
Cochrane tool	√	?							
CEBP	√	?							
Down & Black	√	?						√	√
EPOC	√	?					√		
EPPHP	√	?						√	√
NICE	√	?					√	√	√
SIGN 50	√	?					√	√	√
Wells								√	√
Valentine & Cooper (DIAD)	√	?	√		√		√	√	√

These tools also

- Do not enable consistent classification across the full range of studies (experimental and quasi-experimental)
- Can lead to overly simplistic and inappropriate classification of risk of bias across quasi-experiments
- Are not sufficiently detailed to evaluate the execution of the methods of analysis, including complex statistical procedures
- Often do not evaluate appropriate risk of bias criteria for social experiments

Theory and evidence on QEDs

- Overall, the evidence on international development suggests that high quality quasi-experimental studies based on statistical methods yield comparable results to experiments

- This seems to be particularly the case when evaluators have knowledge about the allocation rule, and so can model it, or the allocation rule is exogenous (Cook, Shadish and Wong, 2008; Hansen et al. 2011)

- But theory and evidence also suggests that if they are not appropriately implemented, these designs can yield misleading results (e.g. Glazerman et al., 2003)

- We should focus on identifying the circumstances under which these approaches yield accurate causal inference

So a tool like MSMS is not appropriate

Exhibit 1: ***The Maryland Scale of Scientific Methods***

A. Research Designs

	Before-After	Control	Multiple Units	Randomization
Methods Score				
Level 1	O	O	X	O
Level 2	X	O	O*	O
Level 3	X	X	O	O
Level 4	X	X	X	O
Level 5	X	X	X	X

Source: Sherman et al. 1998

Internal validity of QEDs

- Impact evaluation assignment based on:
  - Randomised assignment (experimental studies)
  - Exogenous variation (natural experiments and RDDs)
  - Selection by planners and or self-selection by participants

- Risk of bias of quasi-experimental designs largely depends on:
  - The validity of the technique to ensure group equivalence
  - The implementation of the method
  - Other factors such as spillovers/contamination, hawthorne effects

Principles for assessing RoB (Higgins et al. 2011)

- Do not use quality scales
- Focus on internal validity
- Assess risk of bias in results not quality of reporting
- Assessment requires judgement
- Choose domains based on a combination of theoretical and empirical considerations
- Report outcome specific risk of bias

And we would add:

- Assessment should be based on both study design, as well as execution of the design and analysis

- Ideally will provide a common framework for evaluation of risk of bias for different types of designs

RoB tool being developed

- Build in the structure and RoB concept of existing tools including CEBP, EPOC and Cochrane.

- Address the statistical and conceptual assumptions underpinning the validity of quasi-experimental designs based on statistical methods.

- Provide a common framework of 8 evaluation criteria for the assessment of risk of bias and confidence levels (internal validity) in impact evaluation using experimental and the quasi-experimental designs based on statistical methods.

Evaluation criteria	Category of bias	Relevant questions
1. Mechanism of assignment / identification	Selection bias	For experimental designs: is the allocation mechanism appropriate to generate equivalent groups? Does the model of participation capture all relevant observable and unobservable differences in covariates between groups?
2. Group equivalence in implementation of the method	Confounding	Is the method of analysis adequately executed? Are the observable results of the counterfactual identification process convincing? Are all likely relevant confounders taken into account in the analysis? Is the estimation method sensitive to non-random attrition?
3. Hawthorne effects	Motivation bias	Are differences in outcomes across groups influenced by participant motivation as a result of programme implementation and, or monitoring?
4. Spill-overs and cross-overs	Performance bias	Is the programme influencing the outcome of the individuals in the control group (including compensating investments for control groups)?
5. Selective methods of analysis	Analysis reporting bias	Is the method of analysis or specification model used by the author selectively chosen? Is the analysis convincingly reported (and available for replication)?
6. Other sources of bias	Other biases	Are the results of the study subject to other threats to validity (e.g. placebo effects, courtesy bias, survey effects, inadequate

Full internal validity assessment will also take account of

7. Confidence Intervals and significance of the effect	Type I and Type II error.	Is the study subject to a unit of analysis error not adequately accounted for? Is the study subject to heteroscedasticity not accounted for? Is the study not taking into account possible heterogeneity in effects? Are the lack of significant effects driven by a lack of power?

Mechanism of assignment

- *There are other credible methods of identification than randomisation.*
- *However, some study designs (RDDs, longitudinal studies) and methods of analysis (IV) are better able to eliminate or account for unobservables than others (PSM, cross-sectional regression).*
  - *E.g. Miguel and Kremer (2004) Worms ??*

Assumption of assignment mechanisms

Method	Validity assumption
Randomised control trial (RCT)	Assignment /participation rule external to participants and effectively random (observables and unobservables balanced)
Regression discontinuity design (RDD)	
‘Natural experiment’ (instrumental variables)	
Difference-in-differences (DID) regression	Adjusts for time-invariant unobservables; assumes no time-varying unobservables
Propensity score matching (PSM)	Assumes unobservables are correlated with observables

Group equivalence: execution of method of analysis

- *Internal validity relies heavily on the execution of the method of analysis (e.g. group equivalence, efficiency of the participation equation, appropriateness of the instrumental variable);*
- *While a degree of qualitative judgment is required for all methods, it is particularly apparent in QEDs.*

Threats to validity in execution

- RCTs: insufficient observations, non-random attrition
  - E.g. high attrition in Miguel & Kremer (2004) Worms
- DID: attrition, failure to control for time varying covariates in DID regression
  - Farmer field schools eg Orozco-Cirilo, 2008
- PSM: lack of covariate balance
  - DANIDA (2012) covariate imbalance
- RDD: ‘fuzzy’ discontinuity
- IV: exogeneity of instrument
  - Duflo and Pande (2009) ‘Dams’ – instrument not exogenous

Spillovers and other forms of contamination

- Non-geographically clustered designs (when they should be) eg worms, mosquito nets, sanitation etc
- Differential contamination by compensatory intervention (which affects outcome)
- 'Survey effects' (measurement as treatment)
  - E.g. Deworming trials (Taylor Robinson et al. 2012)
  - E.g. WASH and diarrhea studies (Kremer et al. 2009)

Hawthorne & John Henry effects

- *Motivation bias caused by implementation and evaluation monitoring can explain differences in outcomes in intervention studies.*
- *In contrast, expectation (placebo) effects are embodied within the causal mechanisms of many social interventions, so that isolating them may be less relevant (and in many cases is in any case not possible)*
  - *E.g. interventional vs. observational data*
  - *Intensive forms of data collection (cf. WASH and child diarrhoea les)*

Selective methods of analysis

- *The method of analysis used in the IE should be the 'best' (likely least biased) given the available research design*
- *We recognise the importance of theory-based exploratory research tradition in the social sciences. But statistical methods when used inappropriately can provide ample opportunities for biased results-based research*
  - E.g. Pitt and Khandkher (1998) in Bangladesh – construction of outcomes
  - Ricker-Gilbert (2008) in Bangladesh: uses 2-step Heckman model, but presents non-significant IMR coefficient
  - Carlberg (2012) in Ghana: no IV identifiers are significant, and no test presented for exogeneity

Precision of the effect

- Misinterpretation of statistical significance (confidence interval) due to:
- Heterogeneity of effects (eg binomial distribution of outcome)
- Heteroschedasticity and lack of normality
- Unit of analysis errors (allocation at village, analysis at household level)
- Lack of statistical power (small sample size)

Options for 'summary' risk of bias

- Reviewer judgement to determine most important domains to determine high and low risk of bias, preferably specified in protocol
  - E.g. for highly subjective outcomes, blinding more important
  - Some outcomes at less risk of unobservable selection bias than others e.g. time savings resulting from access to amenities – White (2008)
  - In other other cases, unobservable selection bias is particularly problematic (e.g. microfinance due to risk aversion)

- Independent double coding and inter-rater check
- Examine sensitivity of results to risk of bias status

Thank you

Visit:

www.3ieimpact.org

http://www.campbellcollaboration.org/international_development/index.php