October 2018

A practitioners' tool for critical appraisal of randomized assignment and quasi-experimental program evaluations

Annette N. Brown, PhD

Principal Economist
FHI 360

@fhi360research @anbrowndc

fhi360
THE SCIENCE OF IMPROVING LIVES

GEIS 2018

Annette N. Brown

Annette N. Brown, PhD is Principal Economist at FHI 360 and the principal investigator for the ICT4D evidence review project. Brown has worked in international development for more than twenty years in executive, management, and research positions.

@anbrowndc

Hannah J. Skelly

Hannah J. Skelly is a Program Manager for the Mobile Solutions, Technical Assistance and Research (mSTAR) program in the Digital Development Unit at FHI 360. For the past ten years, Skelly has worked with donors, private partners and governments to provide technical support, lead project design and implement programs in education, health and ICT4D.

@hjskelly

fhi360
THE SCIENCE OF IMPROVING LIVES

GEIS 2018

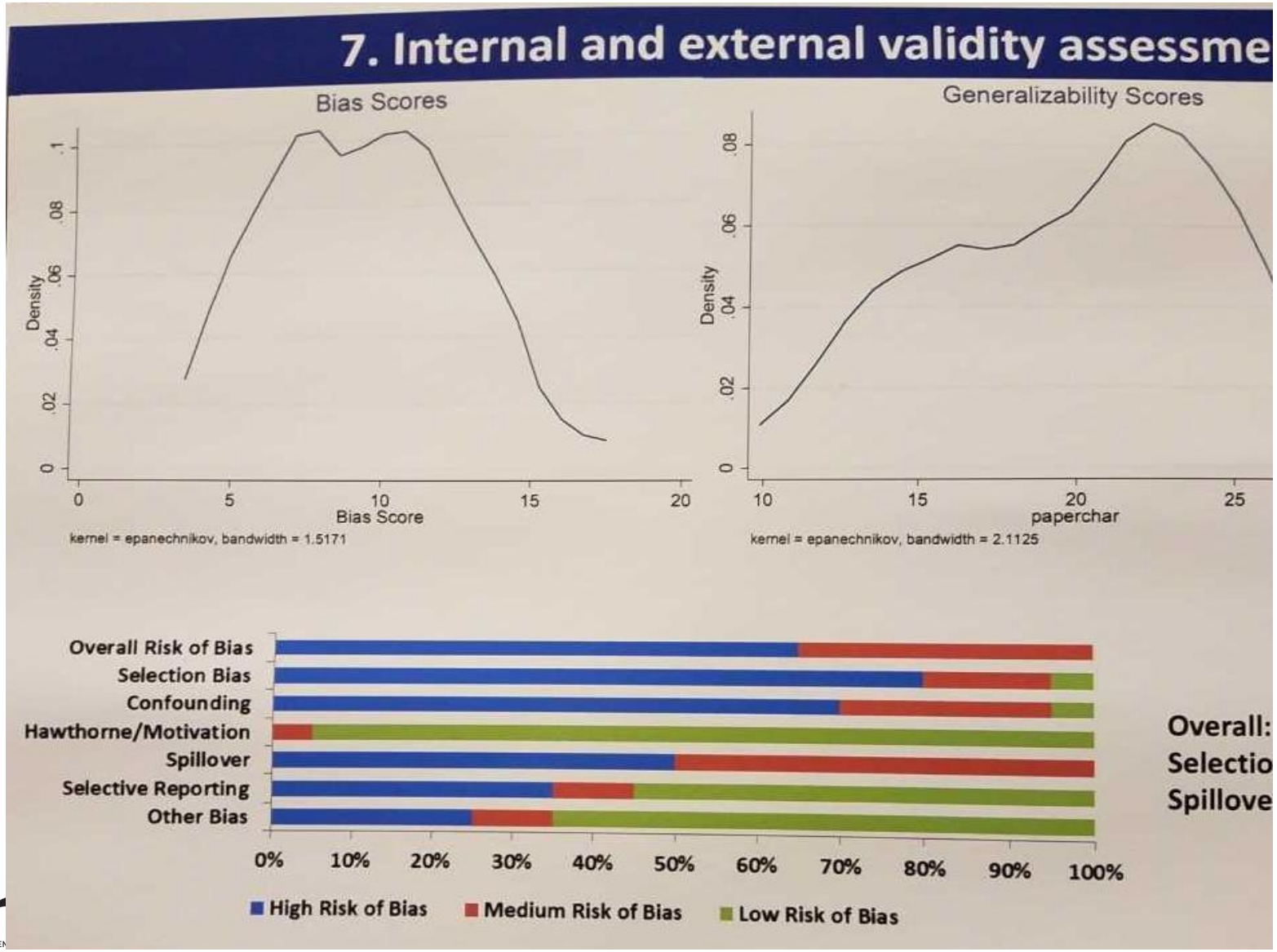
Problem

- Implementers and policy makers often rely on a single or limited number of studies without critical appraisal
- Desire to do critical appraisal in rapid reviews
- Existing critical appraisal tools
  - Narrowly focused
  - Too subjective or arbitrary
  - Too complex

Problem

Study and Outcome	***Assignment mechanism***	***Confounding***	***Selection bias***	***Deviations from intended interventions***	***Performance bias***	***Outcome measure-ment bias***	***Analysis reporting bias***	***Blinded observers***	***Blinded analysts***
Alhassan et al., (2016)									
1-Staff experiences with clients	Yes	Probably No	Probably No	Probably Yes	Yes	Probably Yes	No	Unclear	Unclear
Alhassan et al., (2016)									
2-Staff motivation levels	Yes	Probably No	Probably No	Yes	Yes	Probably No	No	Unclear	Unclear
Alhassan et al., (2015)									
3-Patient safety & risk status	Yes	Probably Yes	Yes	Probably Yes	Yes	Probably Yes	No	No	No
Ananthpur et al. (2014)									
1-Information availability & participation	Probably Yes	Probably No	Probably Yes	Probably Yes	Yes	Probably No	Probably No	No	No
Ananthpur et al. (2014)									
2-Public goods	Probably Yes	Probably Yes	Probably Yes	Probably Yes	Yes	Probably Yes	Probably No	No	No
Banerjee et al. (2014)									
1-Police behaviour (Decoy survey outcomes)	Yes	Probably Yes	Yes	Probably Yes	Yes	Yes	Probably Yes	No	No

Problem


7. Internal and external validity assessme
Bias Scores
Density
0
.02
.04
.06
.08
.1
0
5
10
15
20
Bias Score
kernel = epanechnikov, bandwidth = 1.5171
Generalizability Scores
Density
0
.02
.04
.06
.08
10
15
20
25
paperchar
kernel = epanechnikov, bandwidth = 2.1125
Overall Risk of Bias
Selection Bias
Confounding
Hawthorne/Motivation
Spillover
Selective Reporting
Other Bias
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
High Risk of Bias
Medium Risk of Bias
Low Risk of Bias
Overall:
Selectio
Spillove

Objectives

- Cover both experimental and quasi-experimental designs (but limited to counterfactual-based evaluations)
- Accessible to practitioners with masters level training without requiring advanced statistical understanding
- Focus on internal validity
- Account for research process, including pre-registration, publication status, and internal replication (objective and observable)
- Produce single classification

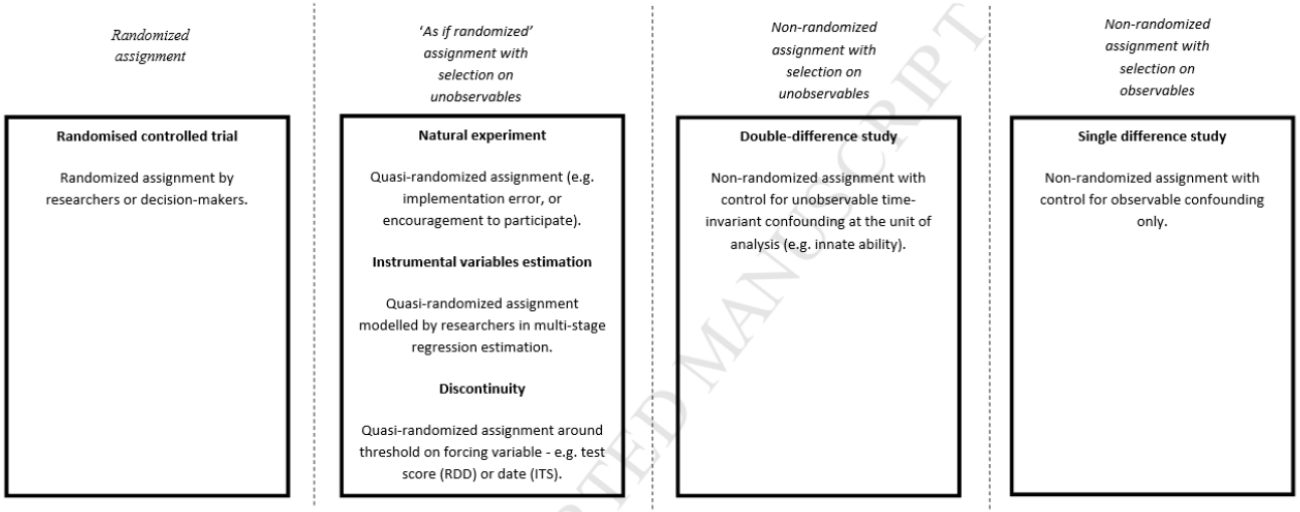
Methods for tool development

- Review of existing critical appraisal tools
- Develop draft tool
- Pilot in context of a rapid review
- Revise the tool
- Pilot in context of a rapid review
- Revise the tool
- Pilot with practitioners and students in the context of a few studies
- Revise?
- Write and submit manuscript

Literature review

Waddington H, Aloe A, Becker BJ, Djimeu EW, Hombrados JG, Tugwell P, Wells G, Reeves B, Quasi-experimental study designs series –Paper 6: Risk of bias assessment, Journal of Clinical Epidemiology (2017), doi: 10.1016/j.jclinepi.2017.02.015.

Figure 1 Study design decision flow for studies of effects using statistical methods



First version features

- Start with Waddington, et al. categories and use as the starting number
- Questions specific to starting category
- Points added based on
  - Objective features based on stage of research
  - Subjective features based on
    - Comparability
    - Outcome measurement
    - Analysis and reporting
- Possibility to subtract a point for robust single difference observational
- Reference to pre-registration within questions
- Collapse four numeric scores into three categories

fhi360
THE SCIENCE OF IMPROVING LIVES

GEIS 2018

First pilot: ICT4D economic growth rapid review

Authors (Year)	Short title	ICT intervention category	Identification strategy	Publication status	Other considerations	Limitations
Aker et al. (2016)	Payment mechanisms and anti-poverty programs	Digital financial services	Random assignment	Journal	Replication +	Few
Ali et al. (2018)	Building fiscal capacity in developing countries	Data systems	Difference in differences	Working paper	None	Many
Banerjee et al. (2017)	E-governance, accountability, and leakage in public programs	E-government	Random assignment	Draft paper	None	Some
Barnwal (2017)	Curbing leakage in public programs	E-government and digital identity	As-if random	Draft paper	None	Some
Batista and Vicente (2013)	Introducing mobile money in rural Mozambique	Data literacy	Random assignment	Working paper	Outcome measurement -	Some
Blumenstock et al. (2015)	Promises and pitfalls of mobile money in Afghanistan	Digital financial services	As-if random	Journal	None	Some
Cadena and Schoar (2011)	Remembering to pay	Digital information services	Random assignment	Working paper	None	Some
Carballo et al. (2016)	The border labyrinth	Data systems	Difference in differences	Working paper	None	Many
Dammert et al. (2015)	Integrating mobile phone technologies into labor-market intermediation	E-government and digital information services	Random assignment	Journal	None	Few

First pilot, ICC results

- Two raters: one researcher and one practitioner

Intraclass correlations
Two-way random-effects model
Absolute agreement

Random effects: **Target** Number of targets = **23**
Random effects: **Judge** Number of raters = **2**

Rating	ICC	[95% Conf.	Interval]
Individual	.7793696	.55276	.8994816
Average	.8760064	.7119709	.9470812

F test that
ICC=**0.00**: F(**22.0, 22.0**) = **8.11** Prob > F = **0.000**

Note: ICCs estimate correlations between individual measurements
and between average measurements made on the same target.

Challenges

- Interpretation of identification strategy, especially determining as-if random designs
- Confusion about publication status
- Tendency for practitioner to judge based on external validity
- Requirement to compare to pre-registration or pre-analysis
- Awkwardness of point system (lower is better, skip patterns in tool)

Revisions

- Develop an Excel template for rating
- Separate pre-registration into its own characteristic
- Make questions about reporting and analysis more direct/objective
- Reverse the numeric order for coding
- Add more explanation for identification strategy
- Revised templates for inquiries to authors

Rating Excel template

If starting value = 3							If starting value = 4								
Not comparable? -1	Concern outcome measure? -1	Concern reporting or robustness? -1	Working/ draft paper or report? -1	Pre-registered and consistent? +1	Public replication? +1		Not comparable? -1	Concern outcome measure? -1	Concern reporting? 1	Working/ draft paper or report? -1	Pre-registration -1 or +1 (or null)	Public replication? +1		Numeric code	Adjectival code

Second pilot, ICC results

Three raters: one researcher, one practitioner, and one intern (recent Masters student)

```
. icc rating target judge

Intraclass correlations
Two-way random-effects model
Absolute agreement

Random effects: target              Number of targets =         17
Random effects: judge               Number of raters  =          3
```

rating	ICC	[95% Conf.	Interval]
Individual	.64811	.3911618	.8387661
Average	.8467525	.658402	.9397827

```
F test that
  ICC=0.00: F(16.0, 32.0) = 6.45                Prob > F = 0.000
```

Challenges

- Interpretation of identification strategies, especially for the new rater (encouragement designs, IV)
- Consistency in coding for working papers and pre-registration, non-response from corresponding author
- Multiple designs in a single paper

Further revisions

- Suggested search string for replication studies
- Additional hints for determining identification strategy
- Instructions to code for the evidence the practitioner wants to use

Next steps

- Pilot the next revision for a limited set of studies with practitioners and masters students, including a short survey
- Explore possibility to test validity against more complex critical appraisal tools

October 2018

Thank you!

Annette N. Brown, PhD

abrown@FHI360.org

@fhi360research @anbrowndc

Cite as: Brown, AN, Skelly, HJ, & Clahane, K. (2018) A practitioners' tool for critical appraisal of experimental and quasi-experimental program evaluations. Manuscript in progress.

fhi360
THE SCIENCE OF IMPROVING LIVES

GEIS 2018