Title registration for a systematic review: Within-study design replications of social and economic interventions: map and systematic review

Hugh Waddington, Paul Fenton Villar, Jeffrey C. Valentine

Submitted to the Coordinating Group of:

☐	Crime and Justice
☐	Education
☐	Disability
☒	International Development
	☐ Nutrition
☐	Social Welfare
☐	Methods
☐	Knowledge Translation and Implementation
☐	Other:

Plans to co-register:

☒	No
☐	Yes ☐ Cochrane ☐ Other
☐	Maybe

Date submitted:
Date revision submitted:
Approval date:

Title of the review

Within-study design replications of social and economic interventions: map and systematic review

Background

Campbell Systematic Reviews are increasingly incorporating non-randomised quantitative causal designs. Konnerup and Kongsted (2012) had estimated that one-half of Campbell reviews incorporated these designs. We estimate that 80 per cent of the reviews published in the Campbell Library between 2012 and 2016 include non-randomised studies in all substantive review groups. The inclusion of non-randomised studies is usually justified by the lack of randomised study evidence for specific interventions, to improve external validity, or to incorporate unintended or adverse effects (e.g. Higgins and Green, 2011).

Non-randomised approaches for causal inference (hereafter, non-randomised studies, NRS) include study designs and methods of analysis which incorporate 'selection on unobservables', including natural experiments, regression discontinuity designs, panel data difference studies and methods such as instrumental variables (Imbens and Wooldridge, 2009). These can be distinguished from NRS with 'selection on observables' which control directly for confounding in adjusted analysis (e.g. statistical matching, analysis of covariance, multivariate regression).

The main problem in ascertaining causal validity in NRS is that the underlying assumptions are often untestable, for example the 'unconfoundedness' assumption in the case of selection on observables. In contrast, unconfoundedness is met by definition in appropriately designed and implemented randomised controlled trials (RCTs). Hence, treatment effects estimated in NRS may differ from the findings of RCTs, over and above sampling error, due to confounding which leads to bias in the effect estimate (internal and statistical conclusion validity). But even when there is no confounding in the effect estimate, treatment effects may differ between NRS and RCT of the same intervention due to differences in the sample in which the treatment effect is estimated (external validity).

An increasingly popular way of validating methods used in NRS is the internal replication study. Internal replication studies compare effect sizes obtained from benchmark experiments (usually RCTs) with those obtained using non-randomised comparison group designs from an overlapping study population. Hence, they are also referred to as 'within-study comparisons' (Glazerman et al., 2003). An early review found that non-randomised

internal replication studies estimated different effect sizes from RCTs (Glazerman et al., 2003). More recent evidence, however, suggests that non-randomised studies in which the method of treatment assignment is known or credibly modelled at baseline, can produce similar findings as the RCTs (Cook et al., 2008; Hansen et al., 2013). In addition, high quality meta-analyses also suggest internally valid non-randomised studies can produce the same pooled effects, although potentially with less precision (Waddington et al., 2017) (e.g. Petrosino et al., 2012; Baird et al., 2013).

Objectives

The purpose of the study is to inform the Campbell Collaboration's approach to incorporating non-randomised studies of effects in systematic reviews of international development interventions. The objectives are as follows:

Review question 1: what is the extent of evidence of non-randomised internal replication studies of experiments of social and economic interventions?

Review question 2: what is the validity of the methods used to undertake non-randomised internal replication studies of social and economic experiments in low- and middle-income countries, and what are the comparative effects and bias coefficients estimated?

Existing reviews

Table 1 lists existing reviews of internal replication studies. Some are of particular literatures, for example labour market studies (Glazerman et al., 2003); others cover particular designs, for example regression discontinuity (Chaplin et al., 2018). Few of the reviews appear to have been conducted systematically. The one review dedicated to evidence from social and economic development programmes in low- and middle-income countries (L&MICs) (Hansen et al., 2013) was not based on systematic approaches to identifying studies, critically appraising them (of either the NRS or the associated RCTs) or statistical synthesis.

Table 1: Examples of reviews of within study comparisons

Authors	*Title*	*Publisher*
Glazerman, Levy and Myers (2003)	Nonexperimental versus Experimental Estimates of Earnings Impacts	The Annals of the American Academy
Cook and Wong (2007)	Empirical Tests of the Validity of the Regression Discontinuity Design	Institute for Policy Research Northwestern University Working Paper Series

Authors	*Title*	*Publisher*
Cook, Shadish and Wong (2008)	Three Conditions under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparison	Journal of Policy Analysis and Management
Shadish and Cook (2009)	The Renaissance of Field Experimentation in Evaluating Interventions	Annual Review of Psychology
Cook and Steiner (2010)	Case Matching and the Reduction of Selection Bias in Quasi-Experiments: The Relative Importance of Pretest Measures of Outcome, of Unreliable Measurement, and of Mode of Data Analysis	Psychological Methods
Hansen, Klejnstrup and Andersen (2013)	A Comparison of Model-Based and Design-Based Impact Evaluations of Interventions in Developing Countries	American Journal of Evaluation
Wong, Valentine and Miller-Bains (2017)	Empirical Performance of Covariates in Education Observational Studies	Methodological Studies
Chaplin et al. (2018)	The Internal and External Validity of the Regression Discontinuity Design: A Meta-Analysis of 15 Within-Study-Comparisons	Journal of Policy Analysis and Management

Intervention

Included studies in the map can be of any social or economic intervention. The causal benchmark and comparison study must also derive from the same intervention and time-period. Studies may have any comparison condition (e.g. no intervention, wait-list, alternate intervention) provided the comparison condition was the same for the benchmark control and replication comparison group.

In addition, included studies in the systematic review must be of social and economic development interventions. Studies excluded from the map and systematic review include those based on clinical and bio-medical interventions (e.g. McKay et al., 1998).

Population

Studies included in the map could be conducted among any population (e.g. low-, middle- and high-income countries; among general programme participants or lab studies conducted among students). The causal benchmark and comparison study must derive from the same

study population so to minimise the possibility of a factor other than study design confounding estimates of bias. In contrast, studies included in the systematic review are limited to those in low- and middle-income countries. We will also address in the analysis issues relating to treatment effect estimates across different study samples (e.g. average treatment effect versus local average treatment effect in regression discontinuity versus average treatment effect on the treated/untreated in statistical matching; intention to treat versus complier average causal effect in instrumental variables).

Outcomes

Studies can be of any outcome variable. We will not exclude studies on outcome and we will collect data on all outcomes reported in included studies.

Study designs

Included studies must report treatment effects for a benchmark causal study, which may be an RCT or a study with quasi-random assignment (e.g. natural experiment), alongside treatment effects for a non-randomly assigned comparison replication or a non-randomly assigned treatment replication. The replicated comparisons could be constructed using any method (e.g. statistical matching, regression discontinuity, difference in differences, adjusted regression estimation etc.). Glazerman et al. (2003, p.65) define a replication study as follows: "researchers estimate a program's impact by using a randomized control group and then re-estimate the impact by using one or more non-randomized comparison groups." As noted above, we broaden the definition of the benchmark study to also include non-randomised treatment groups, and also non-randomised methods with *a priori* low risk of bias (regression discontinuity design and natural experiments using quasi-randomised methods of allocation) (e.g. Somers et al., 2012).

Studies will be excluded that do not use as a causal benchmark study with a priori low risk of bias, or do not incorporate outcomes data collected from human study participants. Studies conducting analyses of artificial or synthetic populations are therefore excluded. In addition, studies will be excluded that use non-experimental techniques to adjust for circumstances where the randomisation process was compromised, or that compared the predicted estimates of ex-ante models or general equilibrium models to estimates from ex-post RCTs.

References

Baird, S., Ferreira, F.H.G., Özler, B. and Woolcock, M., 2013. Relative Effectiveness of Conditional and Unconditional Cash Transfers for Schooling Outcomes in Developing Countries: A Systematic Review. Campbell Systematic Reviews, 2013:8.

Chaplin, D.D., Cook, T.D., Zurovac, J., Coopersmith, J.S., Finucane, M.M., Vollmer, L.N. and Morris, R.E. (2018). The internal and external validity of the regression discontinuity design: a meta-analysis of 15 within-study comparisons. Journal of Policy Analysis and Management. DOI: 10.1002/pam.22051

Cook, T. D., Shadish, W. and Wong, V. 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. Journal of policy analysis and management, 27 (4), 724–750.

Cook, T.D. and Steiner, P.M., 2010. Case Matching and the Reduction of Selection Bias in Quasi-Experiments: The Relative Importance of Pretest Measures of Outcome, of Unreliable Measurement, and of Mode of Data Analysis. Psychological Methods, 15 (1), 56-68.

Cook, T.D. and Wong, V.C., 2008. Empirical Tests of the Validity of the Regression Discontinuity Design: Implications for its Theory and its Use in Research Practice, Annals of Economics and Statistics, GENES, 91-92, 127-150.

Glazerman, S., Levy, D.M. and Myers, D. 2003. Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Science, 589 (1), 63–93.

Hansen, H., Klejnstrup, N.R. and Andersen, O.W., 2013. A Comparison of Model-Based and Design-Based Impact Evaluations of Interventions in Developing Countries. American Journal of Evaluation, 34 (3), 320-338.

Higgins, J.P.T. and Green, S. (editors), 2011. Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011. Available from www.handbook.cochrane.org.

Imbens, G.W. and Wooldridge, J.M., 2009. Recent development in the econometrics of program evaluation. Journal of Economic Literature, 47 (1), 5-86.

Konnerup, M. and Kongsted, H., 2012. Do Cochrane reviews provide a good model for social science? The role of observational studies in systematic reviews. Evidence & policy, 8 (1), 79–96.

McKay, J.R., Alterman, A.I., McLellan, A.T., Boardman, C.R., Mulvaney, F.D. and O'Brien, C.P., 1998. Random versus nonrandom assignment in the evaluation of treatment for cocaine abusers. J Consult Clin Psychol, 66 (4), 697-701.

Petrosino, A., Morgan, C., Fronius, T.A., Tanner-Smith, E.E. and Boruch, R.F., 2012. Interventions in Developing Nations for Improving Primary and Secondary School Enrollment of Children: A Systematic Review. Campbell Systematic Reviews 2012:19 DOI: 10.4073/csr.2012.19

Somers, M.-A., Zhu, P., Jacob, R. and Bloom, H., 2013. The Validity and Precision of the Comparative Interrupted Time Series Design and the Difference-in-Difference Design in Educational Evaluation. MDRC Working Paper on Research Methodology. September 2013.

Waddington, H., Aloe, A.M., Becker, B.J., Djimeu, E.W., Hombrados, J.G., Tugwell, P., Wells, G. and Reeves, B., 2017. Quasi-experimental study designs series – paper 6: risk of bias assessment. Journal of Clinical Epidemiology, 89, 43-52.

Wong, V. C., Valentine, J. C. and Miller-Bains, K., 2017. Empirical performance of covariates in education observational studies. Journal of Research on Educational Effectiveness, 10 (1), 207-236.

Review authors

Lead review author: The lead author is the person who develops and co-ordinates the review team, discusses and assigns roles for individual members of the review team, liaises with the editorial base and takes responsibility for the on-going updates of the review.

Name:	Hugh Waddington
Title:	Senior Evaluation Specialist and Assistant Professor
Affiliation:	3ie and London School of Hygiene and Tropical Medicine
Address:	London International Development Centre
City, State, Province or County:	London
Post code:	WC1H 0PD
Country:	United Kingdom
Phone:	+44 20 7958 8350
Email:	hwaddington@3ieimpact.org

Co-author(s): (There should be at least one co-author)

Name:	Paul Fenton Villar
Title:	Research Student
Affiliation:	University of East Anglia
City, State, Province or County:	Norwich
Country:	United Kingdom
Email:	P.Fenton-Villar@uea.ac.uk

Name:	Jeffrey C Valentine
Title:	Professor and Program Coordinator, Educational Psychology, Measurement, and Evaluation
Affiliation:	Department of Counselling and Human Development, College of Education and Human Development, University of Louisville
City, State, Province or County:	Louisville, KY
Country:	USA
Email:	jeff.valentine@louisville.edu

Roles and responsibilities

Give a brief description of content and methodological expertise within the review team. It is recommended to have at least one person on the review team who has content expertise, at least one person who has methodological expertise and at least one person who has statistical expertise. It is also recommended to have one person with information retrieval expertise.

Please note that this is the *recommended optimal* review team composition.

- Content: Jeff Valentine (JV), Paul Fenton Villar (PFV), Hugh Waddington (HW)
- Systematic review methods: JV, PFV and HW
- Statistical analysis: JV, PFV and HW
- Information retrieval: JV, PFV and HW

Funding

Financial support gratefully acknowledged from the American Institutes for Research under Campbell Methods Grant CMG1.11.

Potential conflicts of interest

No authors have been involved in the development of relevant interventions or primary research, and are not aware of any actual or potential conflicts of interest.

Preliminary timeframe

- Date you plan to submit a draft protocol: August 2018
- Date you plan to submit a draft review: October 2018