Risk of bias assessment in credible quasi-experimental studies[1]

Hugh Waddington, Ariel Aloe, Betsy Becker, Barney Reeves, Peter Tugwell and George A. Wells

Abstract: *Rigorous and transparent critical appraisal is a core component of high quality systematic reviews. Well-conducted quasi-experiments have been empirically shown to estimate credible, unbiased treatment quantities. Conversely, when inappropriately designed or executed, these estimates are likely to be biased. This paper draws on recent advances in risk of bias assessment. It presents an approach to evaluating the internal validity of credible quasi-experiments. These are non-randomised studies using design-based approaches to control for unobservable sources of confounding such as difference studies, instrumental variables, interrupted time series, natural experiments and regression discontinuity designs. Our review suggests that existing risk of bias tools provide, to different degrees, incomplete transparent criteria to assess the validity of credible quasi-experiments. We argue that a tool is needed to assess risk of bias consistently across credible quasi-experiments. Drawing on existing tools, in particular Cochrane's new tool for non-randomized studies of interventions (Sterne et al., 2014), we discuss domains of bias and suggest directions for evaluation questions.*

Keywords: risk of bias, systematic review, meta-analysis, quasi-experiment, natural experiment, instrumental variables, regression discontinuity, interrupted time series, difference in differences, propensity score matching

[1] Thanks are due to David Wilson for suggesting the term 'credible quasi-experiments' at the Campbell Collaboration Methods Group Symposium in Belfast, May 2014, as well to other participants at that meeting and participants the Alliance for Health Systems Research workshop on quasi-experimental studies at Harvard School of Public Health, November 2013.

1. Introduction

Researchers in health and the social sciences quantify statistically valid treatment effects – that is, changes in outcomes which are attributed to a particular treatment – using a range of credible quasi-experimental approaches (Dunning, 2012; Reeves et al., this issue). Quasi-experiments are referred to by various names including natural experiments,[2] observational studies or simply non-randomized studies.[3] 'Credible quasi-experiments' are defined here as approaches that use rigorous designs and methods of analysis which can enable studies to adjust for unobservable sources of confounding. Approaches discussed explicitly in this paper are difference-in-differences, instrumental variables, interrupted time-series, natural experiments, and regression discontinuity designs. Often these designs are combined with methods to control for observable confounding such as statistical matching (e.g. propensity score matching, PSM) and adjusted regression analysis.

All quantitative causal studies are subject to biases relating to design (internal validity) and methods of analysis (statistical conclusion validity) (Shadish et al., 2002). In the same way that experimental studies (randomized controlled trials, RCTs) can have methodological problems in implementation (for example, contagion (contamination), poor allocation concealment, non-random attrition, and so on), inappropriately designed or executed quasi-experiments will not generate good causal evidence. Quasi-experimental studies are, however, potentially at higher risk of bias than their experimental counterparts (Higgins et al., 2012; Rubin, 1974), with perhaps the most critical biases for causal inference being confounding and bias in selection of the reported result. They are also harder to assess than

[2] A UK Medical Research Council (Craig *et al*. 2011) guidelines refers to on quasi-experimental designs as 'natural experiments'; we use 'natural experiment' to refer specifically to designs where exogenous variation in treatment exists, for example due to random errors in treatment targeting and geographical variation.

[3] Randomized studies are defined here as studies in which assignment to the intervention of interest is determined randomly.

RCTs, requiring greater qualitative appraisal of potential biases, which in many cases may need to draw on advanced theoretical and statistical knowledge.

Systematic critical appraisal, operationalized through 'risk of bias' assessment, provides assurance of the credibility of causal studies (Higgins and Green, 2011) and their trustworthiness for decision-making (Chalmers, 2014). Risk of bias tools provide transparency about the judgments made by reviewers when performing assessments. They are usually organized around particular domains of bias, and provide the specific 'signaling questions' which enable reviewers to evaluate the likelihood of bias.

This paper discusses how to operationalize risk of bias assessment for credible quasi-experiments. Section 2 discusses internal validity and Section 3 reviews existing risk of bias tools. Section 4 presents proposed evaluation criteria and signaling questions. Section 5 concludes by proposing an agenda for research in the further development of a risk of bias tool.

2. Internal validity of credible quasi-experiments

Habicht and Victora (1999) distinguish probability evaluation designs, which are able to quantify with statistical precision the change in outcomes attributed to a treatment, from plausibility designs, which attempt to rule out observable confounding but are unable to address important sources of bias, in particular those arising from unobservables.[4] These authors explicitly limit probability evaluations to randomized controlled trials. However, evidence is emerging which suggests quasi-experiments which are able to address

[4] A third category, adequacy evaluations refers to descriptive methods which are not able to address confounding (e.g., uncontrolled pre-test post-test studies where no attempt is made to rule out external factors which may explain observed changes in outcomes) (Habicht and Victora, 1999). These would include single case quasi-experimental designs and other 'small n' approaches.

unobservable confounding can produce the same effect sizes as RCTs in pooled analysis (Table 1).[5,6]

Table 1 Pooled effects of RCTs and credible quasi-experiments

Treatment	Design	Pooled odds ratio[4]	95% confidence interval		P>|z|	Tau-sq	I-Sq	Num obs
Conditional cash transfer (vs control)[1]	RCT	1.43	1.21	1.69	0.000	0.05	90.8%	15
	RCT and QE	1.43	1.28	1.59	0.000	0.04	88.7%	22
Education intervention (vs standard intervention)[2]	RCT	1.33	1.20	1.46	0.000	0.02	90.9%	43
	QE	1.34	1.20	1.52	0.000	0.02	96.7%	16
Microcredit (vs control)[3]	RCT	0.99	0.93	1.05	0.437	0.00	0.0%	4
	QE	0.99	0.88	1.12	0.074	0.00	61.6%	3

Notes: 1/ Baird et al. (2013); outcome is school enrolment. 2/ Petrosino et al. (2012); outcomes is school enrolment and attendance. 3/ Vaessen et al. (2014); outcome is ‘woman makes household spending decisions’. 4/ Pooled odds ratios estimated by inverse-variance weighted random effects meta-analysis. Quasi-experiments (QE) included in the analyses are difference-in-differences, instrumental variables, propensity score matching and regression discontinuity. Source: author calculations based on reported data.

Credible quasi-experiments account for unobservable confounding by design, either through knowledge about the method of allocation or in the methods of analysis used. They are

[5] We note that authors and journal editors may have incentives for selective publishing of favorable comparisons between randomized and non-randomized studies. The examples presented in Table 1 are from systematic reviews (SRs) of socio-economic interventions in low- and middle-income countries supported by the Campbell Collaboration International Development Coordinating Group (IDCG). The findings on experimental and quasi-experimental approaches are representative of the body of evidence in SRs supported by the IDCG. Other examples of comparisons of RCTs and quasi-experiments include Lipsey and Wilson (1993) who provide a meta-analysis of North American social programs and Vist et al. (2009) who compare RCTs and cohort studies in health care studies. Evidence is also available from (within-study) design replication – that is, studies which attempt to compare the same experimental treatment groups with non-randomized comparison groups using quasi-experimental methods. One meta-study suggested significant differences between results from RCTs and quasi-experiments for US and European labor market programs (Glazerman et al., 2003). However, design replications using well-conducted quasi-experimental methods, in which participation has been carefully modelled, have also shown the same results as the RCTs they are replicating (Cook et al., 2008; Hansen et al., 2011).

[6] As noted by Duvendack et al. (2012), effect sizes estimated from credible quasi-experiments may differ empirically from those from RCTs due to differences in external validity – that is, due to the population sampled and the type of treatment effect estimated (see also Aloe et al., this issue).

considered more credible than approaches which rely solely on covariate adjustment of observable confounders (Dunning, 2012; Shadish et al. 2002), whose validity usually relies on unverifiable assumptions about the strength of the correlation between confounding that can be observed and that which cannot (i.e., that observables and unobservables are strongly correlated).

In quasi-experimental designs that use information about the method of allocation to estimate a treatment effect, the ability of the study to identify a causal relationship rests on assumptions that the identifying variables which determine assignment are highly correlated with treatment status but not caused by the outcomes of interest (reverse causality) nor related to any of the other causes of the change in outcome (observable and unobservable confounding) – that is, they are ‘exogenous’. This is the same rationale on which randomized assignment is based, hence we adopt the term ‘as-if randomized’ (Dunning, 2010) for quasi-experimental designs which are, in theory, able to account for all sources of confounding, including unobservables. We differentiate these designs from ‘non-randomized’ quasi-experiments which are only able to account for observable confounding and unobservables under particular conditions (Figure 1).

‘As-if randomized’ quasi-experiments include instrumental variables (IV), interrupted time series (ITS), natural experiments (NE) and regression discontinuity (RD) designs. In natural experiments treatment is assigned ‘as-if randomly’ due to decisions in implementation or take-up, for example by an arbitrary boundary, whether by service provision jurisdiction (Snow and Richardson, 1965) or perhaps according to treatment practice (e.g., Zafar et al., 2014), errors in implementation (e.g., Morris et al., 2004), or manipulated by researchers who are using ‘randomized encouragement’, where participants are exposed randomly to information about an intervention which itself may be universally available (King et al., 2009).[7] We may also include non-random methods of assignment such as alternation here.

'As-if randomized' quasi-experiments also include regression-discontinuity designs (RD) which exploit local variation around a cut-off on an ordinal or continuous 'forcing' variable used by decision-makers to determine treatment.[8] Examples include treatment assignment by diagnostic test score (e.g., Bor et al., 2014), age (e.g., Card et al., 2009), or date as in multiple case-series interrupted time-series (ITS) design.[9]

In the special case of instrumental variables (Zohoori and Savitz, 1997) and related approaches,[10] researchers use such exogenous variables to model treatment decisions in multiple-stage regression (e.g., 2-stage least squares or simultaneous equations maximum likelihood). Exogenous variables used in IV estimation include all of the variables mentioned above, such as randomized assignment or encouragement, differences in implementation across groups (e.g., Wang et al., 2007) and, frequently, geographical factors such as distance[11] (e.g., Newhouse and McClellan, 1998), weather or climate conditions (Lawlor et al., 2006) and topography (e.g., Duflo and Pande, 2008), among many others.[12]

[7] 'Randomized encouragement' designs are classified as natural experiments because the relationship of interest in the study is not usually the pragmatic question about the effect of such encouragement, but rather the mechanistic question about the effect of the intervention in people who are responsive to encouragement. A randomized encouragement study can be analysed conventionally (using intention-to-treat) or using instrumental variables estimation.

[8] Where the forcing variable is not correlated with assignment precisely (e.g. due to performance biases), the design is referred to as a 'fuzzy' RDD and estimation is done using instrumental variables.

[9] ITS here refers to longitudinal panel datasets measured at the disaggregate level (i.e., the same people measured multiple times before and after treatment). It is more common for longitudinal datasets to be clustered at aggregate levels of care (e.g., the health facility or district). In such cases, confounding by secular trends needs to be assessed, for example with reference to a contemporaneous comparison group (controlled interrupted time-series) and an assessment of performance bias.

[10] For example, 'switching regression' models (for a practical example, see Lockshin and Sajaia, 2004). Instrumental variables methods are also used to analyse experimental data, for example to account for non-compliance (Imbens and Angrist, 1994; for an illustration of the approach in epidemiology, see Greenland, 2000).

[11] It is worth noting that location is often endogenous – at least in the long-term, people are able to move to gain access to better services. Hence distance of participant to treatment facility may often not be a good instrument.

[12] See Dunning (2012) for a comprehensive overview of instrumental variables approaches.

Where allocation rules are not exogenous, confounding must be controlled directly in adjusted statistical analyses. Methods such as difference-in-differences (DID, also called double differences, DD), triple differences (DDD) and fixed effects (FE) regression applied to individual level longitudinal data, enable adjustment for time-invariant unobservable confounding (at the level of the unit of analysis) by design, and observable confounding in adjusted analyses.[13] However, these methods are not able to control for time-varying unobservables even in theory. In contrast, single difference (SD) estimation applied to case-control, cohort or cross-sectional data (or in PSM when matching is on baseline characteristics[14]) is not able in theory to control for time-varying or time-invariant unobservables, except in the special case of selection on observables.

In Figure 1, we group these study designs and methods of analysis into three groups ordered from top to bottom according to *a priori* internal validity in addressing confounding. Randomized experiments and 'as-if randomized' design-based quasi-experiments are considered the most credible methods in theory. Non-randomized quasi-experiments are considered less credible in theory, with differencing methods applied to individual level data being favored over non-randomized studies relying solely on analysis of observables.

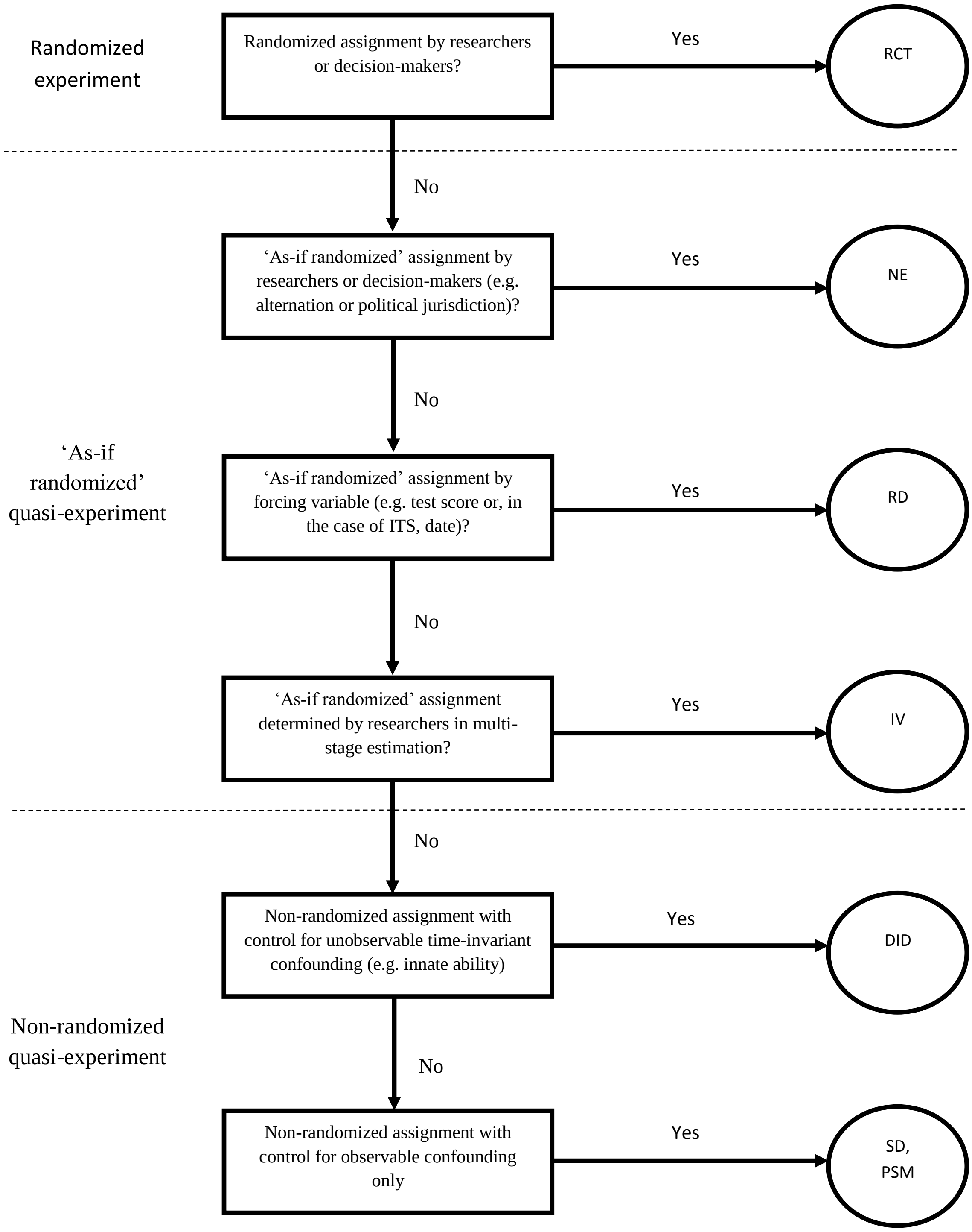
The choice of design for a comparative assessment should capture the information needed to classify a study in this proposed hierarchy (see also Reeves et al., this issue). However, the extent to which designs produce valid causal inferences in practice also depends on the quality of implementation of the approach and the statistical conclusions drawn (see also Vandenbroucke, 1989). Particular flaws in implementation can lead to studies being assessed as being of lower quality than suggested by the *a priori* categories in Figure 1; indeed we

[13] Difference studies can only adjust for unobservable confounding at the unit of analysis, hence it is important to distinguish studies where data analysis is at the individual level, from those where data analysis is conducted at the aggregate level such as the practitioner, health facility, and community or higher.

[14] Difference studies should usually be accompanied by statistical matching (e.g., propensity score matching, PSM) in order to identify the treatment effect among observations in the region of common support (Heckman, 1998).

would expect many quasi-experiments to be downgraded because the assumptions underlying the design are not met.[15]

Figure 1 Study design decision flow for causal studies using statistical methods


Randomized experiment
Randomized assignment by researchers or decision-makers?
Yes
RCT
No
'As-if randomized' assignment by researchers or decision-makers (e.g. alternation or political jurisdiction)?
Yes
NE
No
'As-if randomized' quasi-experiment
'As-if randomized' assignment by forcing variable (e.g. test score or, in the case of ITS, date)?
Yes
RD
No
'As-if randomized' assignment determined by researchers in multi-stage estimation?
Yes
IV
No
Non-randomized assignment with control for unobservable time-invariant confounding (e.g. innate ability)
Yes
DID
Non-randomized quasi-experiment
No
Non-randomized assignment with control for observable confounding only
Yes
SD, PSM

Source: Authors draw on Waddington et al. 2012.

[15] Conversely, strong implementation might in rare cases lead to studies being assessed as of higher quality. An example would be a SD study where selection of participants is based on observable characteristics which are measured at baseline and appropriately modelled in the analysis.

Risk of bias assessment should therefore incorporate design assessment and implementation of analysis (Littell et al., 2008). While the underlying domains of bias (e.g., confounding, biases in outcomes data collection and reporting etc.) are relevant across designs, the criteria used to verify them may differ between 'as-if randomized' and non-randomized study groups and even within groups themselves. For instance, let Z be an exogenous variable determining assignment, T be a dummy variable representing treatment assignment, and Y be the outcome of interest. For 'as-if randomized' studies (NE, IV and RD), the validity assessment will need to incorporate the following criteria: information provided about the relationship between Z and T – in particular, nonzero and monotonic causal relationship between Z and T (Bound et al., 1995); the relationship between Z and Y – that is, Z is not affected by Y or any of its causes and only affects Y through T (exogeneity, also called the 'exclusion restriction'); and the relationship between treated units – Z for one treatment unit does not affect T for another treatment unit (crossovers), T for one treatment unit does not affect Y for another treatment unit (spillovers) and there is no variation in T across treatment units (e.g. due to measurement errors) – collectively referred to as the 'stable unit treatment value assumption' (Chiba, 2010).[16] However, the 'signaling questions' on which each of these propositions can be verified will differ between NE, IV and RD. For example, IV and RD require greater assessment of the statistical methods (e.g., the appropriate bandwidth around forcing variable and functional form specification for RD). The assessment of non-randomized quasi-experiments will be based on different criteria and different signaling questions (Figure 2).

[16] The degree of homogeneity of the relationship between T and Y across individuals induced to treatment by Z is also of interest for external validity (Angrist & Pischke, 2009). See also fn. 6.

Figure 2 Assumptions underpinning validity of credible quasi-experiments

Natural experiments and instrumental variables:

- Relationship between assignment variable and treatment status is monotonic and highly correlated
- Assignment variable does not affect the outcome except through treatment (the 'exclusion restriction')
- Assignment variable is not caused by nor shares causes with the outcome
- Stable unit treatment value assumption (SUTVA) is satisfied
- There is sufficient variation in assignment variable and appropriate regression specification used for instrumental variables

Regression discontinuity and interrupted time series:

- Forcing variable is continuous, or at least ordinal with sufficient values
- Forcing variable is not confounded by other causes of the outcome (e.g. it is not used to determine allocation to another relevant intervention which affects outcome)
- Forcing variable is not anticipated or manipulable by participants
- Forcing variable determines assignment (SUTVA is satisfied)
- Appropriate bandwidth around forcing threshold and regression specification

Difference studies:

- Differencing (or use of fixed effects) controls for all unobservable time-invariant confounding at the level of the unit of analysis
- All observable sources of time-varying confounding are controlled in adjusted analysis and these are correlated with time-varying unobservable confounding
- Differencing controls for unobservable time varying confounding (the equal trends assumption)
- Comparable observations are used across groups (common support)

Sources: Gertler et al. (2012), Hombrados & Waddington (2012), Schochet et al. (2010).

3. Review of critical appraisal tools

A large number of tools exist to facilitate risk of bias assessment of non-randomized causal studies. Drawing on the systematic review by Deeks et al. (2003) and a search of more recent literature, we selected and appraised relevant risk of bias tools according to the extent to which they identified evaluation criteria and signalling questions for credible quasi-experiments as defined here (Table 2). We included tools aiming to assess both randomized and non-randomized studies (Downs & Black, 1998; Cochrane Effective Practice and

Organisation of Care (EPOC), undated[17]; Hombrados & Waddington, 2012; National Institute for Health and Clinical Excellence (NICE), 2009; Reisch, 1989; Sherman et al., 1998; Scottish Intercollegiate Guidelines Network (SIGN), 2011; Valentine & Cooper, 2008; West et al., 2002). We also included tools aiming to appraise only non-randomized studies (Cowley, 1995; Effective Public Health Practice Project (EPHPP), undated; Kim et al., 2013; Sterne et al., 2014; Wells, undated).

Our analysis indicated that existing tools contain evaluation criteria for domains of bias that are relevant to credible quasi-experiments as defined here. However, most of the tools were not designed to assess causal validity of these studies, meaning that the 'signalling questions' on which biases are evaluated were not sufficiently relevant, particularly in the domains of confounding and reporting biases. For example, randomization (sequence generation and allocation concealment) is usually the only method to account for unobservable confounding that is assessed. No single tool fully evaluated the internal or statistical conclusion validity of credible quasi-experimental designs defined here, including the recent tool by Sterne et al. (2014) which was operationalized for designs more commonly used in health research such as cohort and case-control designs. Only one tool addressed instrumental variables design and statistical matching methods (Hombrados & Waddington, 2012) and three tools presented signalling questions for discontinuity designs, of which the most comprehensive was Schochet et al. (2010). Furthermore, most tools which addressed controlled before and after data (e.g., EPOC, n.d.) did not assess the degree to which time-varying unobservables at the unit of analysis (e.g., patient, practitioner or health facility) were controlled using DID methods applied to disaggregate level data. Most tools that aimed to assess experimental and quasi-experimental studies did not enable consistent classification of experimental and quasi-experimental studies, or of different quasi-experimental designs, across similar evaluation

[17] The EPOC tool was developed drawing on the Cochrane risk of bias tool (Higgins et al., 2011).

criteria (e.g., NICE, 2009). One tool (Hombrados & Waddington, 2012) attempted to enable consistent assessment by evaluation criteria, but it was not sufficiently operationalized to capture *a priori* validity according to the design used in the study (see Figure 1).

Table 2 Assessment of experiments and quasi-experiments in existing critical appraisal tools

	Experiment (RCT)	Natural experiment (NE)	Instrumental variables (IV)	Discontinuity design (RDD)	Interrupted time series (ITS)	Difference study (DID)
Cowley (1995)	NA	N	N	N	P	N
Cochrane EPOC (undated)	Y	N	N	N	P	P*
Downs and Black (1998)	Y	N	N	N	N	N
EPHPP (undated)	Y	N	N	N	P	N
Hombrados and Waddington (2012)	Y	P	Y	P	N	Y
Kim et al. (2013)	NA	N	N	N	P	N
NICE (2009)	Y	N	N	N	Y	N
Reisch (1989)	Y	N	N	N	N	N
Schochet et al. (2010)	NA	NA	NA	Y	P	NA
Sterne et al. (2014)	NA	N	N	N	P	P
Valentine and Cooper (2008)	Y	N	N	P	P	P
SIGN (2011)	Y	N	N	N	N	N
Wells (undated)	NA	N	N	N	P	N
West et al. (2002)	Y	N	N	N	N	N

Notes: Y addresses study design and methods of analysis; P partially addresses these; N does not address; NA not applicable. * Includes controlled before and after only.

To take a recent example, the Cochrane Collaboration has recently developed a tool to assess risk of bias in non-randomized studies of interventions (Sterne et al., 2014). That tool uses sensible evaluation criteria to assess risk of bias, with items grouped at pre-intervention stage (baseline confounding and sample selection bias), during intervention (bias in measurement of interventions, e.g. due to problems of implementation fidelity or in recalling treatment status), and after the intervention has started (time-varying confounding, bias due to departures from intended interventions (performance bias), bias due to missing data (e.g. attrition), bias in measurement of outcomes, and bias in selection of the reported result). But it does not distinguish separately the ability of a study to control for observable versus unobservable sources of confounding, so signalling questions to assess the degree of confounding focus solely on methods of observable covariate adjustment. Furthermore, important sources of biases for particular quasi-experiments arising from justification of the design (e.g., exogeneity of an instrumental variable) and the methods of statistical analysis (e.g., bandwidth around a cut-off) are not sufficiently operationalized for credible quasi-experiments, and some included sources of bias may not be relevant (e.g. non-random attrition in a cross-sectional IV study). The concept of the unbiased 'target trial' to which all non-randomized studies should be compared has been useful in getting reviewers from outside of the clinical trials community to think about sources of bias which they may previously have been unaware. However, the point about the target trial being unbiased is quite crucial, as there are instances where trials may be biased in ways which are not applicable to observational studies (e.g. performance bias due to Hawthorne effects, as discussed below).

To summarize, we are not aware of any single tool that sufficiently distinguishes control for (unobservable) confounding by design from control for (observable) confounding in analysis, for non-randomized studies. Each tool addresses some of the potential biases for particular designs, but none provides the specific signalling questions needed to determine whether quasi-experiments are credible enough to recommend using the results in practice or policy. Application of these instruments is therefore likely to lead to inappropriate risk of bias assessment for credible quasi-experiments.

4. Evaluation criteria for credible quasi-experiments

In this section we discuss evaluation criteria and potential signalling questions for credible quasi-experimental studies. Sterne et al. (2014) categorize seven domains of bias which are relevant for non-randomized studies: confounding; sample selection bias; bias due to missing data; bias in measurement of interventions; bias due to departure from intended interventions; bias in measurement of outcomes; and bias in selection of the reported result. These domains form the basis of evaluation criteria that can be used to operationalize risk of bias assessment for credible quasi-experiments. Our discussion focusses on how these domains apply to credible quasi-experiments, recognizing that the categories are also applicable for RCTs.

Confounding refers to the extent to which causality can be attributed to factors determining outcomes other than the intervention. Confounding factors that have been shown to influence outcomes include self-selection and program placement biases (Sterne et al., 2014). Sources of confounding may be observable or unobservable, and time-invariant (identified at baseline) or time-varying. Studies using quasi-experimental approaches need to argue convincingly, and present appropriate results of statistical verification tests, that the identifying variable determining treatment assignment is exogenous to outcomes, and/or that the methods of analysis are able to control for unobservable (time-varying and time-invariant) confounding (see Hombrados & Waddington, 2012). For example, data permitting, it is useful to make assessments of group equivalence at baseline according to observable covariates (Hansen, 2008), under the assumption that these are correlated with unobservables. Factors which may invalidate group equivalence during the process of implementation, such as time-varying confounding, should also be taken into account in estimation.

Sample selection bias occurs where some eligible treatment units or follow-up periods are excluded from data collection or analysis, and this exclusion is correlated with outcome or intervention status. Examples are non-random attrition and censoring of data (e.g., where outcomes data are not available due to mortality).[18] This is particularly important in retrospective studies and studies where baseline data are not available. Assessment is needed

[18] Sterne et al. (2014) refer to this as inception/lead-time and immortal time biases.

of the extent to which the design and methodology account for sample selection biases (e.g., through the use of Heckman error correction). A related domain is *bias due to missing data*, which is a specific source of selection bias due to attrition and incomplete data collection (e.g., on outcomes, treatment status or covariates measured at baseline and after) (Sterne et al., 2014). Biases due to differential attrition are potentially relevant only in prospective studies but biases due to incomplete data collection are relevant for all designs.

Bias in measurement of interventions is not usually considered problematic where information is collected at the time of the intervention from sources not affected by the outcomes (e.g., enumerators). It is particularly problematic where information about treatment status is from participants after implementation who may have an incentive to misreport, or where recalling the intervention (e.g., its dose, frequency, intensity or timing) is difficult (Sterne et al., 2014). This source of bias is most likely to occur in retrospective studies.

Bias due to departures from intended interventions encompass cross-overs, spillovers and implementation fidelity. Cross-overs or switches (including ‘contamination’ of comparison groups) occur where individuals receive a treatment different from that assigned. They are problematic in non-blinded prospective trials (and double-blinded RCTs with an adaptive design where patients cross over if they do not improve sufficiently), as well as designs where the identification strategy relies on a natural experiment, instrumental variable or regression discontinuity (due to SUTVA). Assessment should therefore be made of the extent to which these are accounted for in design or analysis (such as through intention-to-treat or instrumental variables estimation). Spillovers occur when members of the comparison group are exposed to treatment indirectly, through contact with treated individuals, and are potentially problematic for all controlled studies. Cluster-level analysis may be required to ameliorate these sources of bias and/or an assessment of the geographical or social separation of groups may be needed.

Bias in measurement of outcomes due to recall and courtesy biases is potentially problematic in all studies where outcomes data are self-reported. But other forms of motivational bias are only likely to arise in prospective trials. The classic case is the presence of Hawthorne and John Henry effects affecting motivation of participants when they are aware they are part of a trial (particularly when they are subjected to repeated measurement). As another example, ‘survey effects’ may operate whereby groups are sensitized to information that affects outcomes through survey questions and then subjected to repeated measurement (Zwane et al., 2011). Such effects are less likely to affect motivation where data are collected outside of a trial situation with a clear link to an ‘intervention’, and unlikely to be relevant when data are collected at one period of time as in a retrospective cross-sectional only (Hombrados & Waddington, 2012). Blinding is frequently advocated to reduce bias in outcomes measurement. While it may be impossible to prevent participant knowledge of intervention status (especially in evaluations of socio-economic interventions), blinding of outcome assessors and data analysts usually is feasible, though seldom used.

Bias in selection of the reported result corresponds to selective reporting of outcomes (e.g., among multiple possible outcomes collected), selective reporting of sub-groups of participants, or selective reporting of methods of analysis (e.g., where multiple estimation strategies or specifications are used) (Rothstein et al., 2005; Sterne et al., 2014). These types of bias are particularly likely to be prevalent in retrospective evaluations based on observational datasets (e.g., with many IV analyses), but may also arise in prospective studies including RCTs where the method of analysis or outcomes are chosen based on results (e.g., DID). Presence of a study protocol (pre-analysis plan) can help determine the likelihood of bias (although it is recognized that many such studies still do not contain such plans), as can a strong theoretical approach and the assessment of unusual or uncommon methods of analysis.

5. Operationalizing the approach

Higgins et al. (2011) present principles for risk of bias tools for RCTs.[19] We argue that further development of a tool or tools to assess credible quasi-experiments should firstly, aim to build on the bias domains and signaling questions in existing tools used by reviewers, in particular those articulated by Sterne et al. (2014).[20] Second, the tool should address both the conceptual and statistical assumptions underpinning validity. This means that appraisals of, for example, the plausibility of 'as-if randomization' and the exogeneity of identifying variables in the confounding domain will need to be incorporated. The evaluation of quasi-experiments is notoriously more difficult than that of RCTs, relying to a greater extent on what we might call 'qualitative judgment' informed by both advanced statistical and substantive theoretical knowledge. Appraisal by multiple reviewers and inter-rater reliability assessment is therefore crucial (Higgins & Green, 2011).

Third, an integrated assessment tool, covering multiple study designs (RCTs and multiple quasi-experimental approaches), should incorporate *a priori* validity of the designs as well as their execution. Analysis should therefore be based on what is being reported regarding the assumptions of the designs and the methods with which they are addressed (Littell et al., 2008).

[19] We believe five of these principles are applicable to risk of bias of quasi-experiments: focusing on internal validity, choosing bias domains on theoretical and empirical considerations, reporting bias by outcomes, not using quality scales, and requiring judgment in assessments. While it is possible to contact authors to obtain information, focusing on bias in the data as opposed to in the information reported is likely to require replication which is often infeasible for non-trial evidence. Factors relating to motivation of participants (due to observation) are of major concern in trials. For social interventions, expectations (such as placebo effects) may form an important mechanistic component in the process of behavior change.

[20] Several of the authors have received funding from the UK Medical Research Council to extend the Cochrane non-randomized studies risk of bias tool (Sterne et al., 2014) to incorporate design-based quasi-experimental approaches.

Finally, it is likely that some of the signaling questions used to operationalize evaluation of bias will be design-specific, in particular for confounding and reporting biases. For natural experiments and instrumental variables, this will require qualitative appraisal of the exogeneity of the identifying variable or instrument. For instrumental variables the assessment should also incorporate the significance or goodness-of-fit of the first-stage instrumenting equation, the individual significance of the instruments and results of an over-identifying test (Hombrados and Waddington, 2012). For regression discontinuity, the assessment should incorporate whether the forcing variable is continuous, or at least ordinal with sufficient values,[21] the degree to which assignment is exogenous (i.e., not manipulable by participants in response to incentives), comparison of covariate means either side of the cut-off point, and an assessment of appropriate specification (band-width and use of weighting for matches further from the cut-off point) and functional form (e.g. linear or non-linear relationship between forcing variable and outcome). For difference studies, assessments are needed of the unit of analysis at which the differencing or fixed effects occurs (determining whether time-invariant unobervables are 'differenced away' at, e.g., patient, practitioner or health facility level), covariate balance at baseline, adjustment for relevant time-varying covariates, differential attrition, and the existence of equal trends in outcomes before intervention across treatment and comparison groups (an indicator of whether unobservable confounders are changing differentially across group).

As in the case of randomized studies, information needed to inform risk of bias judgments in quasi-experiments must be collected from the studies (see Higgins & Green, 2011, p. 194-197; Sterne et al., 2014). When specific information about these assumptions is unknown,

[21] Schochet et al. (2010) state that ordinal variables should have at least four unique values below the cut-off and four unique values above it.

reviewers may attempt to obtain such information from the primary study authors.[22] The information obtained from risk of bias instruments can be used in a variety of ways (Ioannidis, 2011). Some feasible alternatives are to use this information as part of the inclusion criteria or to use bias information to create moderator variables for a meta-regression. Ahn and Becker (2011) and Herbison et al. (2006) present evidence that meta-analysis should not be weighted by quality scores. As a final point, determining overall risk of bias is complicated because the degree of bias is a latent construct (i.e., a construct that is not directly observable or measureable), but can be useful (see, e.g., Guyatt et al., 2011). While evidence suggests it is not appropriate to determine overall bias using weighted quality scales (Juni et al., 1999), reviewers have shown that it is possible to assess overall bias based on transparent decision criteria (e.g. the reviews reported in Table 1). Others prefer to code separate indicators of particular biases to serve as potential moderator variables.

6. Conclusions

Current tools used by reviewers do not provide the means to evaluate consistently and appropriately the credibility of quasi-experimental studies to address causality. The paper justifies the further development of a comprehensive tool (Sterne et al., 2014), and suggests how it might incorporate quasi-experiments. Authors have received funding from the UK Medical Research Council to undertake this work. The development of a comprehensive tool should be based on several principles. Risk of bias should incorporate *a priori* internal validity information based on the classification of study design and an assessment of the implementation of the approach. The tool could usefully be operationalized to recognize explicitly credible designs like difference studies, instrumental variables, interrupted time

[22] Some reviewers may believe that absence of such information is enough to exclude a study from a review. This should be explicitly stated as part of the inclusion criteria. Where studies are eligible for inclusion by stated design alone, the presence or absence of this information should be incorporated into risk of bias assessment and methods such as meta-regression can be used to explore systematic differences between primary studies that do or do not provide this information.

series, natural experiments and regression discontinuity, and assess them using consistent evaluation criteria across bias domains. It is likely that different signaling questions will be required for different designs, particularly to address confounding and the reporting of appropriate statistical analyses.

7. References

Ahn, S., & Becker, B. J. (2011). Incorporating quality scores in meta-analysis. *Journal of Educational and Behavioral Statistics*, 36 (5), 553-585.

Angrist, J. and Pischke, S., 2009. *Mostly Harmless Econometrics: An empiricist's companion*. Princeton University Press, New Jersey.

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 DOI: 10.4073/csr.2013.8

Bor, J., Moscoe, E., Mutevedzi, P., Newell. M.L. and Bärnighausen, T., 2014. Regression discontinuity designs in epidemiology: causal inference without randomized trials, *Epidemiology*, 25 (5), 729-37.

Bound, J., Jaeger, D. A. and Baker, R., 1995. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association*, 90, 443-450.

Card, D., Dobkin, C. and Maestas, N., 2009. Does Medicare Save Lives? *Quarterly Journal of Economics*, 124 (2), 597-636.

Chalmers, I., 2014. The development of fair tests of treatments, *The Lancet*, 383 (9930), 1713-1714.

Chiba, Y., 2010. Bias analysis of the instrumental variable estimator as an estimator of the average causal effect, *Contemporary Clinical Trials*, 31, 12-17.

Cochrane Effective Practice and Organisation of Care Group (EPOC), undated. Suggested risk of bias criteria for EPOC reviews. Mimeo.

Cook, T., 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.

Cowley, D.E., 1995. Protheses for primary total hip replacement: a critical appraisal of the literature, *International Journal of Technology Assessment in Health Care*, 11 (4), 770-778.

Craig, P., et al., 2011. Using natural experiments to evaluate population health interventions: guidance for producers and users of evidence [online]. London: Medical Research Council. Available from: https://www.mrc.ac.uk/documents/pdf/natural-experiments-to-evaluate-population-health-interventions/ [Accessed 4 November 2015].

Deeks, J., Dinnes, R., D'Amico, R., Sowden, A.J., Sakarovitch, C., Song, F., Petticrew, M. and Altman, D.G., 2003. Evaluating non-randomised intervention studies, *Health Technology Assessment*, 7 (27) (192 pages).

Downs, S. and Black, N., 1998. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions, *Journal of Epidemiological Community Health*, 52.

Duflo, E. and Pande, R., 2007. Dams, *The Quarterly Journal of Economics*, 122 (2), 601-646.

Dunning, T., 2012. *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press, Cambridge.

Dunning, T., 2010. Design-based inference: beyond the pitfalls of regression analysis? In David Collier and Henry Brady (eds), Rethinking Social Inquiry: Diverse Tools, Shared Standards, 2nd Edition. Rowman and Littlefield, Lantham, MD.

Duvendack, M., Hombrados, J., Palmer-Jones, R. and Waddington, H., 2012. Meta-analysis for sophisticaled dummies, *Journal of Development Effectiveness*, 4 (3).

Effective Public Health Practice Project (EPHPP), undated. Quality Assessment Tool for Quantitative Studies. Mimeo.

Gertler, P., Martinez, S., Premand, P. Rawlings, L. and Vermeersch, C. *Impact evaluation in practice*. World Bank, Washington DC.

Glazerman, S., Levy, D. and Myers, D., 2003. Nonexperimental versus experimental estimates of earnings impacts, Annals of the Academy of Political and Social Sciences. Available at: http://www.povertyactionlab.org/doc/non-experimental-vs-experimental-estimates [Accessed 3 December 2015].

Greenland, S., 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29, 722-9.

Guyatt, G.H., Oxman, A.D., Akl, E.A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Murad, Glasziou, P., deBeer, H., Jaeschke, R., Rind, D., Meerpohl, J., Dahm, P. and Schünemann, H.J., 2011. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. *Journal of Clinical Epidemiology*, 64 (4), 383-394.

Habicht, J.-P., Victora, C.G. and Vaughan, J.P., 1999. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *International Journal of Epidemiology*, 28, 10-18.

Hansen, B.B., 2008. Covariate Balance in Simple, Stratified and Clustered Comparative Studies, *Statistical Science*, 23 (2), 219-236.

Hansen, H., Klejntrup, N., and Andersen, O., 2011. A comparison of model-based and design-based impact evaluations of interventions in developing countries, FOI Working Paper 2011/16.

Heckman, J., 1998. Characterizing selection bias using experimental data, *Econometrica*, 66 (5): 1017-1098.

Herbison, Gay-Smith, and Gillespie (2006). Adjustment of meta-analysis on the basis of quality scores should be abandoned, *Journal of Clinical Epidemiology*, 59, 1249-1256.

Higgins J.P.T. and Green S. (eds), 2011. *Cochrane handbook for systematic reviews of interventions,* Version 5.0.0. London, John Wiley and Sons.

Higgins, J.P.T., Altman, D.G., Gøtzsche, P.C., Jüni, P., Moher, D., Oxman, A.D., Savović, J., Schulz, K.F., Weeks, L. and Sterne, J.A.C., 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials, *BMJ* 2011, 343: d5928.

Higgins, J. P. T., Ramsay, C., Reeves, B. C., Deeks, J. J., Shea, B., Valentine, J. C., Tugwell, P. and Wells, G., 2012. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Research Synthesis Methods*, 4 (1), 12-25.

Hombrados, J.G. and Waddington, H., 2012. A tool to assess risk of bias for experiments and quasi-experiments in development research. Mimeo. The International Initiative for Impact Evaluation, New Delhi.

Imbens, G.W. and Angrist, J.D., 1994. Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62 (2), 467-475.

Ioannidis, J., 2011. Meta-research: The art of getting it wrong, *Research Synthesis Methods,* 1 (3-4), 169-184.

Jüni, P., Witschi, A., Bloch, R. and Egger, M., 1999. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*, 282: 1054-60.

Kim, S.Y., Park, J.E., Lee, Y.J., Seo, H.-J., Sheen S.-S., Hahn, S., Jang, B.-H. and Son, H.-J., 2013. Testing a tool for assessing the risk of bias for nonrandomized studies showed moderate reliability and promising validity. *Journal of Clinical Epidemiology*.

King, G., 2009. Public policy for the poor? A randomised assessment of the Mexican universal health insurance programme, *Lancet* 373, 9673, 1447-1454.

Lawlor, D.A., Davey Smith, G., Mitchell, R. and Ebrahim, S., 2006. Adult blood pressure and climate conditions in infancy: a test of the hypothesis that dehydration in infancy is associated with higher adult blood pressure. *American Journal of Epidemiology,* 163, 608.

Lipsey, M.W. and Wilson, D.B., 1993. The efficacy of psychological, educational, and behavioural treatment: confirmation from meta-analysis, *American Psychologist*, 48, 1181-1209.

Littell, J., Corcoran, J., and Pillai, V., 2008. *Systematic reviews and meta-analysis*. New York: Oxford University Press

Lockshin, M. and Sajaia, Z., 2004. Maximum likelihood estimation of endogenous switching regression models, *The Stata Journal*, 3, 282-289.

Morris, S., Olinto, P., Flores, R., Nilson, E. and Figueiró, A., 2004. Conditional Cash Transfers Are Associated with a Small Reduction in the Rate of Weight Gain of Preschool Children in Northeast Brazil, *Journal of Nutrition*, 134 (9), 2336-2341.

National Institute for Health and Clinical Excellence (NICE), 2009. Quality appraisal checklist – quantitative intervention studies. In: *Methods for the development of NICE public health guidance (second edition)*, April 2009, NICE, London.

Newhouse, J.P. and McClellan, M., 1998. Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*, 19, 17-34.

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.

Reeves, B., Waddington, H. and Wells, G. *Journal of Clinical Epidemiology*, this issue.

Reisch, J., Tyson, J. and Mize, S. 1989. Aid to the Evaluation of Therapeutic Studies, *Pediatrics*, 84 (5), November.

Rothstein, H., Sutton, A. and Borenstein, M. (eds), 2005. *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Wiley, London.

Rubin, D., 1974. Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology*, 66 (5), 689.

Shadish, W., Cook, T. and Campbell, D. 2002. *Experimental and Quasi-Experimental designs for Generalized Causal Inference.* BROOKS/COLE CENGAGE Learning.

Schochet P., Cook T., Deke J., Imbens G., Lockwood J.R., Porter J. and Smith J., 2010. Standards for Regression Discontinuity Designs, Mathematica Policy Research Report, Princeton, NJ.

SIGN, 2011. SIGN 50: A guideline developer's handbook. Revised Edition. Scottish Intercollegiate Guidelines Network, Edinburgh. Available at: http://www.sign.ac.uk/pdf/sign50nov2011.pdf [Accessed 1 December 2015].

Snow, J. and Richardson, B.W., 1965. *Snow on Cholera: being a reprint of two papers by John Snow, MD, together with a biographical memoir by B.W. Richardson, and an introduction by Wade Hampton Frost, MD*. Hafner.

Sterne, J.A.C., Higgins, J.P.T. and Reeves, B.C., on behalf of the development group for ACROBAT-NRSI, 2014. A Cochrane risk of bias assessment tool: for non-randomised studies of interventions (ACROBAT-NRSI). Version 1.0.0, 24 September 2014. Available from http://www.riskofbias.info [Accessed 24 September 2014].

Vaessen, J., Rivas, A., Duvendack, M., Palmer-Jones, R., Leeuw, F.L.,Van Gils, G., Lukach, R., Holvoet, N., Bastiaensen, J., Hombrados, J.G. and Waddington, H, 2014. The Effects of Microcredit on Women's Control over Household Spending in Developing Countries: A Systematic review and meta-analysis, *Campbell Systematic Reviews*, 2014: 8 DOI: 10.4073/csr.2014.8

Valentine, J. and Cooper, H., 2008. A Systematic and Transparent Approach for Assessing the Methodological Quality of Intervention Effectiveness Research: The Study Design and Implementation Assessment Device, *Psychological Methods*, 13 (2), 130-149.

Vandenbroucke, J.P., 1989. Is there a hierarchy of methods in clinical research? *Journal of Molecular Medicine*, 67 (10), 515-517.

Vist, G.E., Bryant, D., Somerville, L., Birminghem, T. and Oxman, A.D., 2009. Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. Reprint. *Cochrane Database of Systematic Reviews* 2008, Issue 3. Art. No.: MR000009. DOI: 10.1002/14651858.MR000009.pub4.

Wang, H., Norton, E.C. and Rozier, R.G, 2007. Effects of the state children's health insurance program on access to dental care and use of dental services. *Health Serv Res*, 42, 1544-63.

Wells, G., undated, Newcastle-Ottawa quality assessment scale. Mimeo.

West, S., King, V., Carey, T.S., Lohr, K.N., McKoy, N., Sutton, S.F. and Lux, L., 2002. Systems to rate of strength of scientific evidence. Evidence Report/Technology Assessment Number 47. AHRQ Publication No. 02-E016.

Zafar, S.N., Libuit, L., Hashmi, Z.G., Hughes, K., Greene, W.R., Cornwell III, E.E., Haider, A.H., Fullum, T.M. and Tran, D.D., 2015. The Sleepy Surgeon: Does Night time Surgery for

Trauma Affect Mortality Outcomes? *The American Journal of Surgery*, doi: 10.1016/j.amjsurg.2014.12.015.

Zohoori, N. and Savitz, D.A., 1997. Econometric approaches to epidemiologic data: relating endogeneity and unobserved heterogeneity to confounding. *Ann Epidemiol*, 7, 251-257.

Zwane, A.P., Zinman, J., van Dusen, E., Pariente, W., Null, C., Miguel, E., Kremer, M., Karlan, D., Hornbeck, R., Gine, X., Duflo, E., Devoto, F., Crepon, B. and Banerjee, A., 2011. Being surveyed can change later behavior and related parameter estimates, *Proceedings of the National Academy of Sciences of the United States of America*, 108 (5), 1821-1826.