On minimizing the risk of bias in randomized controlled trials in economics

Alex Eble, Peter Boone, and Diana Elbourne

Abstract

Estimation of empirical relationships is prone to bias. Economists have carefully identified and addressed sources of bias in structural and quasi-experimental approaches, but the randomized control trial (RCT) has only recently begun to receive such scrutiny. In this paper, we argue that several lessons from medicine, derived from analysis of thousands of RCTs conducted over the past 60 years and establishing a clear link between certain practices and biased effect estimates, can be used to reduce the risk of bias in economics RCTs. We first identify the subset of these lessons applicable to RCTs in economics. We then use them to assess the risk of bias in estimates from economics RCTs published between 2001 and 2011. In comparison to medical studies, we find most economics studies do not report important details on study design necessary to assess risk of bias. Many report practices that suggest risk of bias, though this does not necessarily mean bias resulted. We conclude with suggestions on how to remedy these issues.

* Eble: Brown University and Effective Intervention, Mailing address: Brown University Department of Economics, 64 Waterman Street, Providence, RI 02912, USA (email: alexander_eble@brown.edu) Boone: Effective Intervention, mailing address: Effective Intervention, Centre for Economic Performance, London School of Economics, Houghton Street, London, WC2A 2AE, UK, (email: pb@effint.org) Elbourne: London School of Hygiene and Tropical Medicine, mailing address: Medical Statistics Department, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK (email: diana.elbourne@lshtm.ac.uk) This paper was previously circulated under the title "Risk and Evidence of Bias in Randomized Controlled Trials in Economics". The authors would like to thank Simon Johnson and Miranda Mugford for helpful conversations and Samuel Brown, Garret Christensen, Steven Durlauf, Morgan Hardy, Vernon Henderson, Paul Musgrave, Gareth Olds, Anja Sautmann, Tim Squires, David Weil, Hyunjoo Yang, and participants at NEUDC 2012, the PAA 2013 annual conference, Royal Economic Society 2013 annual conference, Brown University micro lunch and Georgetown University Quantitative Models seminar for many helpful comments, as well as generous input from the editor and three anonymous referees. We thank Effective Intervention for financial support. Eble gratefully acknowledges the financial support of the US National Science Foundation. All remaining errors are our own.

I. Introduction

The practice of assigning different courses of action to different groups and comparing outcomes dates back thousands of years. In the Old Testament, King Nebuchadnezzar orders a group of his subjects to eat rich meat and drink wine while another group is made to adhere to vegetarianism in order to evaluate the merits of the two diets (1 Daniel 11-16, New International Version). Versions of this approach have since been used in countless other efforts to evaluate competing hypotheses, from 18th century studies of scurvy treatment to the A/B testing now common in technology firms.

One particular version of this approach is the Randomized Controlled Trial (RCT). An RCT is usually a large-scale study, prospectively designed to test a small set of hypotheses by randomly assigning treatment(s) to participants. Medical scientists have implemented hundreds of thousands of RCTs since the mid-1940s to test hypotheses about treatment options and inform care decisions[1]. In the 1980s, several studies showed that RCTs in medical research yielded less biased treatment effect estimates than observational studies. The method has been adopted in several scientific fields, as well as by the US Food and Drug Administration (FDA) and other government agencies, as the "gold standard" of empirical evidence (Vader 1998).

Not all RCTs are created equal, however. Meta-analysis of thousands of medical RCTs has revealed several pitfalls that skew effect estimates and lead to erroneous conclusions (Jüni, Altman, and Egger 2001). Broader conclusions based on flawed studies have led to the use of drugs or procedures that bring no measurable benefit and, in some cases, even cause harm (Goldacre 2014). In the past two decades, medical researchers have synthesized this research linking certain design choices and biased results to develop standards for the design and reporting of RCTs. These standards are intended to reduce the risk of bias from the identified pitfalls. Adherence to them is now required for publication in most major medical journals (Plint et al. 2006).

Since the early 2000s, economists and other social scientists have increasingly used RCTs to evaluate hundreds of questions of both academic and policy interest (Parker 2010; Vivalt 2015). Within academia, the RCT is now widely considered a part of the economist's empirical toolkit

[1] There were 183,991 trial protocols registered in the US Government's Clinical Trials database (www.clinicaltrials.gov) as of 11 February 2015. It is perhaps impossible to know how many trials have ever been conducted, as public registration of trial protocols was not common practice until the late 1990's.

(Angrist and Pischke 2010); beyond academia, RCTs are often used to determine government policy as well as guide decisions in large international organizations (Parker 2010; Council of Economic Advisors 2014).

In this paper, we argue that several lessons from the medical literature's work linking pitfalls in trial design and reporting to bias in effect estimates can be used to improve the accuracy of estimates generated by RCTs in economics and other social sciences. The medical literature has spent decades scrutinizing these issues and its potential evidence base spans tens of thousands of already-conducted RCTs. The result of this work is a large body of research establishing a direct link from aspects of RCT design and reporting to biased effect estimates. While there are concerns in the medical literature that may be inappropriate for economics (e.g. strict protocols for blinding participants), several others are already central in empirical economic research: selection bias, non-classical measurement error, attrition, attenuation, and selective reporting. Recent work in economics has highlighted how some of these issues could lead to biased effect estimates (Bruhn and McKenzie 2009; Deaton 2010; Brodeur et al. 2013), but there is no consensus among economists on how an RCT should be designed and reported to avoid these problems (Miguel et al. 2014).

In the first part of the paper, we aim to help fill this gap. We draw upon the medical literature described above and the economics literature on RCTs and program evaluation to identify potential sources of bias in economics RCT estimates for which there is applicable evidence from medicine.

Having made the case for the importance of these issues, the second part of the paper addresses our main research question: have economists running RCTs taken the necessary steps to avoid the relevant bias-inducing pitfalls identified by the medical literature? To answer this question, we attempt to collect the universe of RCTs in economics published between 2001 and 2011 in a set of well-regarded journals. For each study, we then generate two assessments: first, whether the article provides the reader with enough information to evaluate the risk of bias in its estimates, and second, whether the study falls into any of the traps that have been associated with biased treatment effect estimates in medicine.

We find that most studies in our sample do not report several pieces of important information necessary for the reader to determine the risk of bias in the presented results. It is important to note that even in medicine, bad reporting is often associated with poor practice, it may also be the case that bad reporting masks good practice and may not necessarily imply bad

methods (Soares et al. 2004). Several of the studies in our sample report quite well in most regards and, as economics lacks standards for reporting, it is not surprising that reporting of RCTs in economics is uneven. Still, this reporting gap begs for remedy: we argue that the burden of proof of the unbiasedness of a study's results rests with the study's authors.

Among those studies that do report key design and analysis decisions, we find that many fall into precisely the same pitfalls that have biased medical RCTs in the past. Our findings raise concerns about the strength of the conclusions reached in several of the studies scrutinized.

Together, the first and second parts of our study suggest that a core set of reporting and design practices drawn from the medical literature can be used to enhance the accuracy and precision of estimates generated by RCTs in economics. We conclude the paper with a series of suggestions on how to improve RCT design and reporting going forward.

II. Identifying sources of bias in RCTs

Empirical work in economics has made increasing use of the RCT to test theory and generate parameter estimates, yet efforts within economics to address the risk of bias in RCT estimates are limited in scope. Bruhn and McKenzie (2009) show that randomization procedures are often not reported in RCTs and, particularly for small studies, certain procedures are more likely than others to lead to chance imbalances between treatment groups that in some cases cannot be addressed by ex-post adjustments. Franco et al. (2014) find that significant results were 40 percentage points more likely to be published than null results among a set of 221 National Science Foundation-funded studies in the social sciences spanning political science, economics and psychology. Brodeur et al. (2013) and Vivalt (2015) find some evidence of selective reporting in economics RCTs, though much less than in observational studies. Allcott (2015) shows that choice of implementing partner can bias RCT results relative to the expressed treatment effect of interest. Miguel et al. (2014) argue for reporting standards in social science RCTs and document recent efforts to meet this need. These studies are all quite recent and focus on individual aspects of the larger set of biases that threaten RCT results. They are necessarily limited by the relatively small number of RCTs that have been conducted in economics to date.

In this section, we argue that several lessons from the long history of conducting and scrutinizing RCTs in medical research are applicable to RCTs in economics. Though RCTs have long been seen as the "gold standard" of evidence in medicine, a series of studies demonstrated a

negative relationship between methodological quality of medical RCTs and measured effect size. A landmark 1995 article linked problems in trial design to exaggeration of treatment effect estimates (Schulz et al. 1995). Its results have since been confirmed by several other meta-analyses linking certain design and reporting practices to biased estimates (Jüni, Altman, and Egger 2001; Gluud 2006; Dwan et al. 2008; Lesley Wood et al. 2008). These findings instigated a movement to improve and standardize methods of reporting and scrutinizing RCTs.

In the late 1990s, two groups began working independently on a set of reporting standards for use in publication of randomized trials. Their combined efforts resulted in two main outputs. The first is the CONSORT Statement (henceforth "CONSORT"), a set of guidelines for publication of reports of randomized controlled trials. Adherence to these standards is now required by most editors of major medical journals (Schulz, Altman, and Moher 2010). The second is the Cochrane Collaboration, an international organization that facilitates systematic review and meta-analysis of published studies in order to draw overall conclusions about efficacy of various treatments. It publishes a handbook that guides authors on how to conduct these reviews. The handbook includes a section on how to evaluate the risk of bias in estimates generated by RCTs based on the body of literature linking certain trial design and reporting decisions to biased treatment effect estimates. The handbook is updated frequently and has been used in 8,600 systematic reviews of trials[2], which have together assessed the risk of bias in hundreds of thousands of scholarly articles (The Cochrane Collaboration 2010). This increased scrutiny during peer review and after publication has resulted in a reduction, over time, in the presence of the biases described above in medical RCT reports (Plint et al. 2006).

The US Food and Drug Administration (FDA) uses a similar set of standards to approve the sale of pharmaceuticals for public consumption. The progress of studies through each stage of this approval process illustrates the importance of such standards in screening false-positive results. For a drug to be approved by the FDA, it must pass three "phases" of trial. There is increasing scrutiny at each phase, such that phase two trials have a higher burden of proof than phase one but less than phase three, whose standards most resemble the CONSORT standards. Among trials that enter phase two, only 70 percent progress to phase three. Of those, 40 percent fail to show positive results in the two phase three trials usually necessary for FDA approval (Danzon, Nicholson, and Pereira 2005).

[2] According to http://www.cochranelibrary.com/cochrane-database-of-systematic-reviews/index.html, accessed 11 February 2015.

Both CONSORT and Cochrane identify six types of problem associated with systematic bias in treatment effect estimates: selection, performance, detection, attrition, reporting and sample size biases (Jüni et al. 1999; Higgins, Green, and Cochrane Collaboration 2008; Moher et al. 2010). All of these have been treated in the broad economics literature. Selection, attrition, reporting and sample size issues have been dealt with extensively (Ashenfelter, Harmon, and Oosterbeek 1999; Wooldridge 2010). Much of performance and detection biases can be seen through the lens of the Hawthorne effect and non-classical measurement error, respectively, also well-known threats to economists (Duflo, Glennerster, and Kremer 2007).

The medical literature's extensive evidence base, developed over six decades of running RCTs, is what allows it to make a novel contribution to the study of bias in economics RCTs. The CONSORT and Cochrane documents synthesize the results of thousands of studies and hundreds of meta-analyses to pinpoint the most likely candidates for RCT-specific bias and outline practices in avoiding them. These are the lessons we hope to use to improve RCT estimates in economics.

Next, we discuss the sources of bias identified in decades of scrutiny of medical RCTs which we believe are applicable to economics RCTs. For each of the six biases (selection bias, attrition bias, performance bias, detection bias, reporting bias, sample size bias), we first explain the concern and its relation to economics. We then describe the reporting and design criteria that form the basis of the analysis we conduct in Section III.

Sources of bias

Selection bias refers to the concern that systematic differences exist between treatment groups at the outset of the trial that could confound treatment effect estimation. There is a long literature in economics on selection bias in program evaluation, summarized by a recent *Handbook of Labor Economics* chapter (DiNardo and Lee 2011) and also discussed extensively in a *Handbook of Development Economics* chapter on running RCTs (Duflo, Glennerster, and Kremer 2007). The medical literature contributes evidence linking a set of mechanisms through which the RCT-specific process of enrolling participants and assigning them to treatment and control groups can artificially generate a difference between the two unrelated to the treatment effect of interest.

Identified problems with selection bias arise from two main sources. The first is non-random assignment to groups. Historically, participants in medical RCTs have often tried to tamper with or predict the randomization procedure. In other cases, researchers used a randomization

method that led to systematic baseline differences between the two groups. A review of several meta-analyses found that studies with problematic randomization procedures generated results that were 12 percent more likely to be positive than studies with unbiased randomization procedures (Gluud 2006).

The relevant lesson is that it is important for the study to clearly state how randomization was done so that "the reader can assess the methods used to generate the random allocation sequence and the likelihood of bias in group assignment" (Schulz, Altman, and Moher 2010). The Cochrane Handbook echoes this concern:

> *The starting point for an unbiased intervention study is the use of a mechanism that ensures that the same sorts of participants receive each intervention...If future assignments can be anticipated, either by predicting them or by knowing them, then selection bias can arise due to the selective enrolment and non-enrolment of participants into a study in the light of the upcoming intervention assignment.*

Economists and medical researchers have identified another potential pitfall in this category: that systematic differences arise between the stated population from which the sample is drawn and the participants who are ultimately randomized or analyzed (selection bias after entry or attrition bias). Manski (2013) discusses this problem in the context of drug trials run for FDA approval. If the participants of the trial are not representative of the population that the RCT attests to study, he argues, then the resultant treatment effect estimate will be a biased estimate of the population treatment effect[3]. A study that evaluates a smoking cessation drug using only light smokers as participants, for example, is likely to generate a biased estimate of the effect of the drug if it attests to study efficacy for the population of all smokers. Frijters, Kong, and Liu (2015) show evidence of this effect biasing the result of an RCT in rural China. A common issue in non-blinded cluster randomized controlled trials, which are frequent in the economics literature, is how to specify eligibility criteria for the population that will be analyzed. For example, in an intervention providing materials to schools, there is a risk that some parents will switch their children to the better-equipped schools from control or non-study schools. This can lead to biased estimates of the effect of the intervention. In situations like this, the study design can include measures which will reduce or eliminate such bias (e.g. by enumerating children for analysis prior

[3] Note that this is separate from concerns of generalizability. While generalizability deals with the applicability of a treatment effect estimate to a population different from that which generated the estimate, the concern we discuss here is about the internal validity of the treatment effect estimate for the stated population.

to randomization, and/or, agreeing restrictions on school transfers with school authorities), but special care needs to be taken.

To assess adequacy of reporting and risk related to selection bias, we look for three pieces of information. The first is detail about how randomization was performed and, if this is information is present, whether it was done in a way to prevent the two problems associated with randomization discussed above: one, the risk of a bad rule which could itself generate bias, and two, the risk of people predicting or switching the group they are assigned to. We ask if the authors mention the method of randomization (e.g. by computer, stratified, public lottery) or any other information to suggest that a non-deterministic, tamper-proof rule was used to assign individuals or clusters to treatment and control groups.

The second piece of information we look for is detail on who is screened for eligibility, who is eligible, who is enrolled in the trial and who is excluded. . This information is necessary to determine whether, as in Manski (2013), there exists a discrepancy between the putative population being studied and the population for whom the treatment effect is actually estimated. We also used this information to examine whether, due to the nature of the trial design, members of the population included in the primary analysis might have had an opportunity to enter the trial, or switch arms in the trial, post randomization. Where issues could be present, we checked whether the authors attempted to address, or at least reported and/or discussed those issues. Finally, we look to see whether the authors provide a table showing baseline covariates by treatment group which might suggest successful randomization. It is important to note that there even with secure randomization there may be imbalance by chance especially if a trial is small, but also that a potentially problematic allocation sequence could lead to issues of bias even if there was balance on observables, as the selection-on-unobservables literature points out (Manski 2013; Oster 2013).

Attrition bias refers to a systematic loss of participants over the course of a trial, differentially between the trial arms, in a manner that potentially destroys the comparability of treatment groups obtained by randomization. Economists have dealt with attrition thoroughly in the empirical literature on the use of observational data (Heckman 1979; DiNardo and Lee 2011). In the context of an RCT, loss of participants stems from similar reasons: drop-out, missing data, refusal to respond, death, or any exclusion rules applied after randomization. The issue, as in Heckman (1979), is that the incidence of attrition may be partly driven by the treatment group one is in. One famous case from medicine is a study which initially showed a large positive impact of a drug to treat heart disease. The first publication excluded participants who died during the trial,

though mortality differed substantially between control and intervention groups. Subsequent analyses that included all participants according to randomization status, performed by a third party after the initial publication, failed to reject the null of no treatment effect (Temple and Pledger 1980).

Attrition bias can also stem from decisions of whom to exclude from the final analysis. This relates to the decision whether to present analysis according to the "intent-to-treat" (ITT) principle or a "treatment-on-the-treated" (TOT) analysis (also termed per-protocol analysis), the difference between which is well understood in economics as well as in medicine (Duflo, Glennerster, and Kremer 2007). The relevant lesson from medicine is primarily about reporting – the reader should know whether the analysis presented is the ITT or TOT estimate to ensure that an unbiased account of the result of the trial is given.

In our assessment of attrition bias, we look for a few key pieces of information. The first is a clear discussion of how participants flowed through the trial, from enrollment to the final analysis. The relevant lesson from medicine is that it is essential to know how many people drop out in each treatment group, their characteristics, and whether or not this drop-out destroys the balance obtained at baseline through randomization.

The second concern is the application of the "intent-to-treat" principle. We look either for an explicit mention of the principle or, in the absence of its explicit mention, evidence of deviation from it in the main analyses. Specifically, if ITT is not mentioned, we check to see whether the number of participants randomized is equal to the number of participants included in the final analysis and, if there is a difference, whether it is explained. A study is judged to be reported inadequately only if it does not mention ITT, either adherence to it or explaining the reason for and ways in which the study deviated from it, and does not explain discrepancies between the number of participants randomized and the number included in the analysis of outcomes. It is considered to be at high risk of bias if there are substantial unexplained discrepancies between these two figures, or the exclusions described by the authors are likely to introduce bias between treatment and control groups not present at baseline.

Performance bias is also known as the set of "Hawthorne" and "John Henry" or "research participation" effects. There is a documented tendency both in economics and medicine for participants to change their behavior or responses to questions because they are aware of being in a study and, specifically in a trial, are aware of their treatment allocation (Leonard and Masatu 2006; McCambridge, J et al. 2014; Noseworthy et al. 1994; Zwane et al. 2011). This can skew

treatment effect estimates either upwards or downwards. In medicine, blinding of participants is often used to minimize this type of bias. In many economics studies and some medical studies, however, blinding is either ethically or logistically infeasible. For example, in the study of village-level education interventions, blinding participants with a placebo intervention would be unethical (although analysis could be conducted blind to allocation). In some economics studies, blinding may even be contrary to the goals of the research.

The relevant lesson from medicine is that extra scrutiny must be applied in two cases. The first case is when outcomes are subjective (e.g. self-reports of pain or personal opinions). A meta-analysis of studies of acupuncture treatment on back pain showed that while acupuncture was superior to control interventions in unblinded studies, it could not be proven to be superior to sham-interventions in blinded studies (Ernst and White 1998). Though all outcome assessments can be influenced by lack of blinding, there is greater risk of bias with more subjective outcomes. Lack of blinding was associated with a 30% exaggeration in treatment effect estimates in a meta-analysis of studies with subjective outcomes (L. Wood et al. 2008).

The second case is when patients are likely to change their behavior given their knowledge of which group they are assigned to. Knowledge of allocation status has been known to induce some control group participants to seek extra care, which, if effective, would introduce a systematic downward bias on treatment effect estimates. In economics this is often the stated purpose of the research, as in Akresh et al. (2013). In studies attempting to evaluate the effect a specific treatment, e.g. the effect of a medicine on an illness, however, unaccounted-for differential care seeking by treatment group could bias effect estimates.

In our assessment, we look for information on these two concerns when blinding participants to which treatment group they are in is impossible. The first concern is whether the outcomes are subjective enough to be vulnerable to the Hawthorne Effect. The second is whether individuals are aware of the treatment under study and their assignment to treatment or control. If so, we ask whether this might induce them to act in a way that would offset or intensify the impact of the treatment the researchers are intending to measure. We flag this as a concern only when there is likely offsetting/intensifying behavior, such as differential care seeking, not accounted for in the description of the study.

Detection bias (also called assessment bias) is concerned with data collectors unduly influencing either the behavior of participants or the data collected in a way that generates artificial differences between treatment groups. This is likely to work through one of two channels. The first

channel is similar to the placebo effect. CONSORT notes how data collectors' knowledge of the treatment status of each participant may lead them to unconsciously filter the data they collect: "unblinded data collectors may differentially assess outcomes (such as frequency or timing), repeat measurements of abnormal findings, or provide encouragement during performance testing. Unblinded outcome adjudicators may differentially assess subjective outcomes" (Moher, Schulz, and Altman 2001). In a trial in which ill patients performed a walking test with and without encouragement from the data collector, encouragement alone was shown to improve time and distance walked by around 15 percent (Guyatt et al. 1984) and similar impacts of detection bias have been found in other medical RCTs (Noseworthy et al. 1994). The second channel is a simple case of incentive alignment. If data collectors are employed by the organization whose intervention is being evaluated in an RCT, there is a clear conflict of interest that raises concerns about the accuracy of the data collected.

In our assessment, we first look to see whether data collectors are blinded to the treatment status of participants. If the data collectors are not blinded, we then look to see whether the data collectors are contractually related or otherwise linked to the organization administering the treatment in a way which might induce them to bias the data they collect. We also ask whether there is any other reason to suspect data collection might differ between the two arms in a substantive way, such as data collected at different scheduled times or by different individuals for treatment and control groups.

Reporting bias points to the fact that it is exceedingly difficult, in any reading of empirical analysis, to know whether authors are presenting the entirety of the results of the study or only that subset of outcomes which is deemed interesting or sympathetic to the case they are trying to make. Recent meta-analysis has shown evidence of this among studies in economics (Brodeur et al. 2013) and in medicine, the latter of which finds that "statistically significant outcomes had a higher odds of being fully reported compared to non-significant outcomes (range of odds ratios: 2.2 to 4.7)" (Dwan et al. 2008). A meta-analysis of medical studies on anthelminth therapy and treatment for incontinence found that "more outcomes had been measured than were reported." This study calculated that with a change in the assumptions about which outcomes the largest study chose to report, "the conclusions could easily be reversed" (Hutton and Williamson 2000).

To combat this problem, many medical journals require that a protocol and statistical analysis plan be registered with a third-party database before the study begins. These documents record the plan for conduct of the trial, the intended sample size, and the analyses that the

researchers plan to undertake at the end. This is called a "pre-analysis plan" in economics. While there are tools in economics which can help mitigate some types of the multiple comparison problem (Kling, Liebman, and Katz 2007), a recent study in economics demonstrates how separate and contradictory erroneous conclusions could have been drawn from a randomized experiment in Sierra Leone in the absence of a pre-analysis plan (Casey, Glennerster, and Miguel 2012). We acknowledge that pre-analysis plans involve important tradeoffs in the context of economics research (Olken 2015), but argue that, at the very least, the decision of whether or not to have one should be documented in the final publication so that the readers can judge for themselves about the study-specific risk of bias this entails.

Furthermore, to prevent authors from running analyses *ad infinitum* and unduly weighting only those which are statistically significant, medical journals require that both the protocol and subsequent article report which outcome is "primary" and thus given highest credence. For non-primary outcomes, additional labels of "secondary" (pre-planned, but not the primary analysis) and "exploratory" (conceived of after the data was collected and examined) are assigned to the remaining presented results. Though exploratory analyses are seen as informative, they are given less weight than pre-specified analyses, as there is a wealth of evidence of false-positive results from *ad hoc* analyses conducted with the benefit of being able to look at the data first (Oxman and Guyatt 1992; Yusuf et al. 1991; Assmann et al. 2000; Casey, Glennerster, and Miguel 2012).

The sophisticated statistical and econometric tools often employed in robustness checks and sensitivity analysis in economics provide some protection against this risk, and recent work in economics shows that reporting bias may be less of a concern in RCTs than observational studies (Brodeur et al. 2013). Vivalt (2015) also tests for reporting bias in a large set of trials and impact evaluations, finding little evidence of reporting bias in published RCTs.

These studies, however, do not provide enough evidence to evaluate the broader risk of reporting bias in RCTs in economics. Brodeur et al. (2013) limit their analysis of experiments to only 37 articles from three top journals, two thirds of which are non-randomized laboratory experiments, not RCTs. The small sample size and journal spectrum of this exercise limit its generalizability. Vivalt (2015) scrutinizes a larger number of studies than is covered in our paper but focuses on generalizing from impact evaluations in development, which is a substantially different aim than that of our analysis. Commenting on her results related to reporting bias, she also notes that while "these figures look much better than the typical ones in the literature," her

choice of which estimates to use in each eligible paper was "designed…partially to minimize bias, which could help explain the difference."

We look for a series of indicators to inform our assessment of the risk of reporting bias. The first is presence of a pre-registered protocol and/or analysis plan. We realize this is unlikely for many economics studies, particularly those published in our time frame; however the goal of our analysis is to document what is reported in published RCTs in economics and to assess the risk of bias in these studies. The medical literature clearly links over-weighting *post hoc* outcomes to risk of bias (Assmann et al. 2000). The potential for this bias is also documented in the economics literature (Casey, Glennerster, and Miguel 2012). The second piece of information we look for is specification of a "primary" analysis or outcome (in medicine, a "primary endpoint", which is usually one single measure, although study designs can incorporate more than one primary endpoint and clearly specify how they will address multiple testing issues). We recognize this is similarly strict, however we point again to the unambiguous link between the lack of reporting constraints and the likelihood of finding significant results in the medical literature.

Finally, under reporting bias, we examine the interpretation of results. Here we look for a clear and objective description of the study which

- Summarizes the findings of the study
- Considers alternative mechanisms and explanations of the results
- Offers a comparison with relevant findings from other studies and a brief summary of the implications of the study in the context of other outcomes and evidence, evidence which is not limited to evidence that supports the results of the current trial
- Offers some limitations of the present study
- Exercises special care when evaluating multiple comparisons

These five issues, taken directly from CONSORT, set a fairly low bar for what should be reported in the interpretation of a study. We include them in our assessment to determine whether the study expresses irrational exuberance about its results, another form of reporting bias identified in the medical and economics literatures (Deaton 2010), perhaps the result of labeling the RCT as the "gold standard" of evidence.

Sample size bias is better known among economists as the twin concerns of attenuation and undue bias from outliers. An insufficiently large sample size does not in itself lead to biased estimates of the treatment effect, but it can lead to imprecise estimation and, if not properly interpreted, incorrect conclusions (Wooldridge 2010). Sample size calculations should be included in any pre-analysis plan in order to understand the effect size the study is capable of measuring. CONSORT describes the risk of small sample sizes:

> *Reports of studies with small samples frequently include the erroneous conclusion that the intervention groups do not differ, when in fact too few patients were studied to make such a claim. Reviews of published trials have consistently found that a high proportion of trials have low power to detect clinically meaningful treatment effects. In reality, small but clinically meaningful true differences are much more likely than large differences to exist, but large trials are required to detect them.*

Guyatt and Mills and Elbourne (2008) debated the value of small trials in the medical literature, and a recent study of the issue also finds that trials with inadequate power have a high false-negative error rate and are implicated as a source of publication bias (Dwan et al. 2008).

The second concern is that without enough observations, draws from the extreme right or left tail are unduly weighted and could lead to exaggerated results. Two other studies in medicine link small sample sizes to overstating effect size because of the heightened influence of outliers (Moore, Gavaghan, et al. 1998; Moore, Tramèr, et al. 1998). To guard against these problems, both CONSORT and Cochrane expect researchers to conduct sample size calculations before collecting any data and report these calculations in trial publications[4].

In our assessment, we look for a description of the sample size calculation used to design the study in the paper, in a publicly available pre-study registration, or in an online appendix. It is important to note that the reader cannot always infer the necessary sample size from the reported standard errors on an RCT's treatment estimates, as these too are sample moments which are more subject to bias the smaller the sample size is. The inclusion of a prior sample size calculation tells the reader what the trial was designed to measure and allows the reader to see whether there were enough observations collected to test the original hypothesis. It also links the main outcomes

[4] One reader pointed out that our bias assessment tool includes many items (such as sample size calculations) which could be considered "common sense" to include in an RCT report. This emphasizes our points that 1) the absence of much of the information we are looking for is somewhat surprising, and 2) the shortcomings in reporting we identify prevent the reader from determining the risk of bias in many RCTs in economics.

presented to the original design of the trial, which helps guard against specification searching and misrepresentation of ad-hoc analysis.

III. Assessing reporting and the risk of bias in RCTs in economics

Using the issues identified in the previous section, we next attempt to answer two research questions. First, are the recent reports of RCTs in economics providing readers with sufficient information to assess the risk of bias in the study? Second, among these studies, what is the risk of each of the six types of bias, given the empirical evidence linking certain design and reporting choices and exaggerated treatment effect estimates in the medical literature?

Identifying relevant concerns

We first read the literature from economics and medicine on sources of bias in RCT estimates and program evaluation to identify the subset of concerns from the medical literature most applicable to economics RCTs. These concerns are described in the previous section. We then developed a reporting and bias assessment tool to determine, for each study, what is reported and the risk of bias for each identified concern[5]. Next, we attempted to collect all economics articles reporting RCTs published between 2001 and 2011 in a set of 52 major peer reviewed journals. This collection process is described in further detail below. To evaluate the validity of our assessment tool and to provide a benchmark for our assessments of articles in economics, we randomly selected an equal number of articles from three top peer-reviewed journals in medicine. Finally, we applied our assessment tool to both sets of articles.

The assessment tool was designed to facilitate and collect assessments of adequacy of reporting and risk of bias in terms of the six biases discussed above. Following the concerns outlined earlier, there are 12 specific issues we assess spread across the six biases[6] with leading questions to aid assessment. For example: "does the paper give the number of participants in each group included in the analysis, and whether this analysis is according to the "Intention to Treat" principle? If not, is there evidence that the principle was followed?"

The task of the assessor is to make two assessments for each issue: first, does the paper report adequately on the matter, providing the reader with enough information to assess the risk of

[5] The assessment tool is given in the Web Appendix.

[6] We began with 13 and removed one as it was excessively stringent. Details are given in the Web Appendix Where is this?

bias, and second, is the paper at low risk of bias from the relevant threat? The assessor circles either a yes or a no for each question and, if possible, provides a page number and/or explanation in the comment and quote boxes to the right of the question to justify each assessment. We decided on the following rule for assessment of risk of bias: if a paper did not report adequately on the issue, it could not be assessed as having a low risk of bias. This decision reflects our judgment, mentioned earlier, that the burden of proof of the unbiasedness of a study's results rests with the author. The landmark meta-analysis assessing study quality in medicine uses a similar rule (Schulz et al. 1995). We present results on reporting and risk of bias for each individual issue as well as aggregated to the bias level under a simple rule – if a study is inadequately reported or not at low risk of bias for one issue, it is inadequately reported or not at low risk of bias for the relevant bias. We do not create an overall study-level assessment[7], as expectations on both the sign and magnitude of bias vary across issues.

We selected studies for assessment using the following process:

1) We searched EconLit for journal articles published between 2000 and 2009 that contained either the word randomized or randomization (or their alternative UK spellings) in the title or abstract. A search conducted on July 6th, 2010 generated 527 results. This was amended on September 5th, 2012, to expand the time range to include papers from 2010 and 2011. The amendment yielded 235 additional results[8].
2) Within these results, we further limited eligibility by two criteria:

a. We included only articles reporting results of prospectively randomized studies. As we are evaluating study design, it would be inappropriate to include studies not designed as trials (e.g. natural experiments).
b. To limit heterogeneity of study quality, we further restricted eligibility to articles published in the top 50 journals as rated by journal impact within economics, taken from a Boston Fed working paper which ranks economics journals (Kodrzycki and Yu 2006). In the 2012 search amendment, we added papers from the *American Economic Journal: Applied Economics* and the

[7] Several meta-analyses of the risk of bias in medicine follow this practice as well (Spiegelhalter and Best 2003).

[8] We recognize that this is not the universe of published RCTs but believe it is a good approximation. Scanning the abstracts of all articles in these journals published over the period would have been prohibitively time-consuming. Including the word "experiment" in the search terms raises the number of initial results well into the thousands.

American Economic Journal: Economic Policy, from the journals' inception in 2009 onward, in light of their prestige and the volume of RCT reports they publish.

In total, this yielded 54 articles published between 2001 and 2011.

We then conducted a search to collect studies reporting RCTs in three top peer-reviewed medical journals for assessment. This served two purposes – one, to calibrate our assessment tool[9], and two, to provide a benchmark for how enforced standards might improve reporting. Articles in medicine were drawn from the top three medical journals according to impact factor in general and internal medicine on July 6th, 2010 from *Thompson Journal and Citation Reports* (Thompson Reuters 2010). These were *The Lancet, The Journal of the American Medical Association,* and *The New England Journal of Medicine*. This restriction was made for ease of processing, as it reduced the number of eligible studies in each year from several thousand to approximately 350, and to ensure we were evaluating the "gold standard" in medicine as described above. The selection process for medical articles was as follows:

1) We searched Pubmed (a database similar to Econlit indexing articles in medical journals) for all articles reporting clinical trials in the three journals in years for which there was also an eligible economics article (all years in our range save 2002).
2) From this list, we then randomly selected as many articles in a given year as there were eligible articles in economics from that year. Among studies published in a given year, selection was performed by assigning each article a random number between 0 and 1 using a random number generator. We sorted the articles by their randomly assigned number and, beginning with the lowest random numbers, we then selected the required number of articles.
3) We excluded phase one and phase two trials in medicine as their methods, goals and sample size considerations are significantly different from phase three trials, which, similar to the economics trials we are concerned with, are more often used to inform policy.

[9] Given that the medical trials we collected were published in journals that required adherence to the standards in the CONSORT Statement, if we were to find most medical trials were at high risk of many biases (low risk of all biases), we would be concerned that the instrument was too strict (lenient).

The final list of both sets of papers is given in the Appendix. If a trial generated more than one eligible publication, the article published earliest was selected and the remaining associated articles were used to provide additional information for assessment of the main article.

The assessment tool was first piloted by all three authors and Miranda Mugford. Once it was finalized, two authors (AE/PB) first read each article and assessed the adequacy of reporting and risk of bias using the assessment tool individually. For each article, we then discussed our assessments. Any disagreements were resolved through deliberation, the result of which is the final assessment of each study. We adopted this method of individual assessment followed by deliberation for two reasons. First, the exercise was a novel one and we expected our assessments to improve through discussion. Second, we followed the example of several meta-analyses in the medical literature, which find that while independent assessment potentially provides better internal validity of the tool, the rate of agreement between assessors in such processes is often low (Clark et al. 1999). In practice, our mean rate of agreement on an issue was greater than 85 percent.

Results

For four of the six biases in our assessment tool, less than 30 percent of the articles collected are assessed as reporting adequately, and for no type of bias are more than three quarters of the economics articles assessed as reporting adequately. Among the subset of articles in which reporting is assessed as adequate, there are many cases in which there is high risk of bias, that is, in which the authors report having made trial design decisions which are known to have biased estimates in medicine. In the exercise used to calibrate our instrument, we found that medical RCTs, published in journals which require these standards be followed, have substantially better reporting and lower risk of bias, though for none of our bias categories do 100 percent of the articles report adequately or have low risk of bias.

These overall performance ratings mask substantial heterogeneity on the different issues within the six biases. While in some issues (reporting and sample size) few papers are assessed as having low risk of bias, in others (performance and detection) most relevant issues are usually addressed. Indeed, in some cases the papers published in economics that we examine fare no differently than those we examine which are published in the top three medical journals.

Below, we show summary statistics of our assessments at the issue and bias-level, and describe our assessments for each issue in detail. Figure 1 shows simple bar charts with 95 percent confidence intervals documenting performance of economics articles and medical articles in terms

of adequacy of reporting and risk of bias for each of the six biases. Similar charts breaking down the assessments of each bias by issue are given in Appendix 3. Table 1 provides the number of papers assessed as adequately reporting and at low risk of bias at the issue and bias levels with a chi-square test for equality of proportions between the assessments for economics and medicine.

[Insert Table 1 Here]

[Insert Figure 1 Here]

Only 12 of the 54 eligible economics articles (22%) passed all of the reporting criteria for *selection bias*, while 40 of the 54 eligible medical articles did so. Performance varied across the three issues in this bias. Thirty-four of the 54 economics papers reported adequately on their randomization procedure, but five of these used clearly deterministic methods to assign treatment. An alphabetic rule was used in one case and sorting by date of employment commencement was used in another. Less than half of the economics studies provided adequate information about the flow of potential participants in the trial. In the majority of economics articles, information on the number of participants at three important stages - screening for eligibility and exclusion from the study before and after eligibility was assessed - was not given, raising concerns about potential undocumented discrepancies between the declared population of interest and the sample studied (Manski 2013; Frijters, Kong, and Liu 2015). All but six of the 54 economics papers provided a table showing whether there was balance on observables at the time of randomization, suggesting that randomization was usually successful. Two papers that gave this information showed evidence suggesting that the randomization did not achieve the desired balance.

The largest issue related to *attrition bias* was failure to report how many participants progressed through the trial from enrolment to inclusion in the final analysis. More than two thirds of the economics RCTs we assessed had striking inconsistencies between the number of participants they enrolled and the number of observations included in the final analyses which were not discussed in the body of the paper or in the appendixes. The number of observations varied among final analyses in many of these papers, in some cases by up to 30 percent, often with no explanation for the difference. As reported in Table 1, the papers with flow of participants data clearly outlined avoided these problems. We suspect the discipline of monitoring and reporting the flow of participant data encourages trial designers to limit attrition, as well as helping ensure that

authors explain cases of substantial attrition. Reporting of adherence to or deviation from the intent-to-treat principle was adequate in more than half of the studies we assessed. Two of these reported deliberate exclusions that suggested risk of bias.

Thirty-eight of the 54 economics papers reported adequately in terms of *performance bias* and only one of these reported a design decision which raised concerns about risk of bias. In this case, there was possibility of unaccounted-for alternative care-seeking as a result of knowledge of treatment status which could have biased the effect the authors were trying to measure. In the sixteen studies assessed as not reporting adequately, the most common concern was a subjective outcome assessed without blinding and without mention of the possibility of bias from the Hawthorne Effect. These circumstances are linked clearly in medicine to exaggerated treatment effects (Lesley Wood et al. 2008). Overall, assessment of reporting and risk of performance bias in economics articles was not statistically distinguishable from our assessment of medical articles.

Thirty-seven of the 54 studies reported adequately on the issues surrounding *detection bias*. Two of these 37 documented problematic practices. In both cases, the authors explicitly mentioned using data collectors who were employed by the same organization which administered the intervention. Of the seventeen not assessed as reporting adequately, most neglected to say who collected the data, leaving doubt as to whether a similar conflict of interest could have biased the results.

No economics paper was assessed as adequately reporting in terms of *reporting bias*, and therefore none could be assessed as having low risk of bias in this category. This assessment attests to the absence of either a pre-analysis plan or registration of a study protocol prior to implementation of the trial. No economics paper in our sample mentioned either, though we are aware that writing a protocol and registering it is increasingly common in economics. Economics RCT protocol registries have been established by both the American Economic Association and J-PAL, among others.

The other relevant concern is the specification of a primary outcome and the differentiation between planned and ad-hoc secondary analyses. We enthusiastically support, and ourselves practice, conducting analyses conceived after a trial finishes. We agree with the medical literature, however, that they should be described as such to allow the reader to weight the different types of evidence provided in the paper. The final issue in our assessment of reporting bias in economics was interpretation of results. Nearly half of the economics papers did not mention whether there were any limitations in their methods nor did they condition their interpretation of the strength of

their results in light of the many comparisons that they presented. Interestingly, the medical papers in our sample also fared rather poorly in this final regard.

Only two economics papers attested to perform a prior sample size calculation. We are almost certain that some others did (Banerjee et al. 2007; Parker 2010), but as none were reported, overall the economics literature did not report adequately on this bias. We decided against soliciting such information from authors in light of evidence that doing so was likely to lead to biased responses (Haahr and Hróbjartsson 2006) and our rule tying inadequacy of reporting to risk of bias was applied.

We calculated subgroup-specific bias assessments for a few categories of interest for both economics and medical RCTs. These results are shown in Figures 2-4. We found that more recent studies in economics (i.e. from the 2010-2011 amendment to our initial search) performed similarly to their earlier-published counterparts (Figure 2), though we suspect this is improving with the establishment of trial registries and the increased attention these issues have received in the past few years. In medicine, we observe better reporting and lower risk of the six biases in the more recently published group, likely a consequence of the increasing use of CONSORT guidelines by journal editors.

[Insert Figures 2-3 Here]

Papers reporting the results of economics RCTs taking place in developing countries (Figure 3) had more issues with performance, detection, and attrition bias than papers reporting the results of trials taking place in the US, Canada, and Europe. Among economics studies taking place in the developing world, data collectors were more often related to the intervention being applied, outcomes were more often subjective, and the number of observations was less stable among the final analyses within a paper. We find no such differences between those medical RCTs run in developed countries compared to those run in developing countries.

[Insert Figure 4 Here]

The performance of papers published in the "top five" journals (*Econometrica*, the *American Economic Review*, the *Journal of Political Economy*, the *Quarterly Journal of Economics* and the *Review of Economic Studies*) was similar to performance of papers in the other 47 economics journals we included for all six of the biases (Figure 4).

IV. Ways Forward

We have presented evidence that a large proportion of RCTs in economics published between 2001 and 2011 did not report many pieces of information necessary for the reader to assess the risk of bias in the evidence provided. Among those studies that do report this information, we found that several made many of the same design choices that have been shown to lead to biased results in medical RCTs. As a result, we conclude that these trials are at unnecessarily high risk of presenting exaggerated treatment effect estimates.

The economics literature has begun to address several of these issues. A series of "toolkits" on how to conduct RCTs have been put forth (Duflo, Glennerster, and Kremer 2007; Glennerster and Takavarasha 2013) and groups such as the Berkeley Initiative for Transparency in the Social Sciences conduct annual meetings which focus heavily on improving methods and transparency in social science research.

Our paper contributes novel evidence to this discussion. We make the case that a series of lessons from the medical literature is applicable to economics RCTs and use them to scrutinize RCTs published in economics journals between 2001 and 2011. We show that there is ample room for these lessons to be used to improve both the reporting and design of RCTs in economics.

To ensure that the evidence from RCTs published in the economics literature is as reliable as possible, we echo calls elsewhere (Miguel et al. 2014) to establish a system of reporting standards for RCTs in economics, similar to the CONSORT guidelines widely accepted in the medical literature. The contents of such a system would have to come from a consensus among economists on what constitutes good practice as well as which data are necessary to assess risk of bias. This should draw on the recent toolkits mentioned above.

As Miguel et al. (2014) note, some standards for trials in economics will necessarily differ from those in medicine. The medical standards are imperfect by their own admission and, as discussed earlier, the goals of some economics research are in direct conflict with certain CONSORT strictures. A good starting point for the departure from medical reporting standards is the admissibility of and weight placed on non-pre-specified outcomes, given the sophisticated statistical and econometric tools often employed in robustness checks and sensitivity analysis.

However, in many areas the "good reporting" requirements for economics trials and medical trials will be similar. The CONSORT guidelines included in our bias assessment tool were suitable for all of the economics studies we examined in this paper. They address most

situations (multiple endpoints, non-blinded, cluster randomization) typically found in economics trials which are less common in medical trials (Campbell et. al. 2012). In cases where the guidelines were implemented by authors, such as including "Flow of participant" diagrams, we noted a substantial lowering of specific risk of bias in economics papers.

We strongly suggest that, at the very least, the following issues from CONSORT be part of any set of guidelines for RCT design and reporting: a CONSORT-style diagram of flow of participants; requiring either registration of protocols/pre-analysis plans prior to randomization or a discussion of why this was decided against; requiring pre-specification of a primary outcome accompanied by a link to the relevant sample size calculation conducted prior to trial entry; and, in cases where appropriate, insistence on the intent-to-treat principle for the primary analysis.

There are a few productive avenues of inquiry we leave to future research. Monte-Carlo simulation of the impacts of different types of bias using existing data from economics RCTs and censuses could illustrate the likely magnitude of the biases outlined here. Standards on reporting related to generalizability, discussed elsewhere (Vivalt 2015; Allcott 2015), are arguably of similar importance and there is a rich literature on how to assess this in reports of RCTs (Rothwell 2006).

Lastly, we would like to mention that a major weakness of our study is the number of assessors we used. Our assessment task was a long and tedious one and almost certainly not without some human error. An increase in the number of evaluators for each paper would almost certainly improve the reliability of our results. Nonetheless, our independent initial assessment by multiple individuals follows best practice in systematic review and the high level of agreement in our independent assessments suggests a high degree of objectivity. The application of our assessment tool to ongoing research would shed additional light on how recent efforts to improve the quality of economics RCTs have fared.

V. Conclusion

In this study, we make two main contributions. First, we identify a series of lessons from the medical literature on sources of bias in RCT estimates that are applicable to economics RCTs. Second, we use these lessons to assess the adequacy of reporting and risk of six major biases in economics RCTs published in 52 top economics journals between 2001 and 2011. We find that these articles often do not provide the reader with essential information on design and reporting decisions related to identified sources of bias. We conclude that RCTs in economics are at a far higher risk of reporting exaggerated treatment effects than necessary given what we know from

medicine about how to minimize bias in RCTs. We finished by suggesting, as have others, that one means by which to minimize this risk would be for economists to develop and adopt a set of reporting guidelines to ensure clarity and precision in the reports of RCTs. We offered several suggestions for the content of such standards.

Going forward, we hope that our study will contribute to the establishment and acceptance of a set of standards for designing and reporting RCTs. Such standards would serve two purposes. First, they would improve the quality of RCTs going forward. Second, they would serve as a tool to help scholars and policymakers in assessing the risk of bias in estimates from existing studies. The medical example has shown that such repeated scrutiny is likely to increase efforts by researchers themselves to avoid these pitfalls in the design, execution, and analysis of their trials. We strongly believe that these efforts would lead to higher quality evidence and, we hope, improve the usefulness of RCTs in learning and policy decisions.

REFERENCES

Akresh, Richard, De Walque, Damien, and Harounan Kazianga. 2013. *Cash Transfers and Child Schooling: Evidence from a Randomized Evaluation of the Role of Conditionality*. SSRN Scholarly Paper ID 2208344. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2208344.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117-1165.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30. doi:10.1257/jep.24.2.3.

Ashenfelter, Orley, Colm Harmon, and Hessel Oosterbeek. 1999. "A Review of Estimates of the Schooling/earnings Relationship, with Tests for Publication Bias." *Labour Economics* 6 (4): 453–70.

Assmann, Susan F., Stuart J. Pocock, Laura E. Enos, and Linda E. Kasten. 2000. "Subgroup Analysis and Other (mis) Uses of Baseline Data in Clinical Trials." *The Lancet* 355 (9209): 1064–69.

Banerjee, A., R. Banerji, E. Duflo, R. Glennerster, D. Kenniston, S. Khemani, and M. Shotland. 2007. "Can Information Campaigns Raise Awareness and Local Participation in Primary Education?" *Economic and Political Weekly*, 1365–72.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. 2013. "Star Wars: The Empirics Strike Back." *Discussion Paper Series, Forschungsinstitut Zur Zukunft Der Arbeit* 7268.

Bruhn, Miriam, and David McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1 (4): 200–232.

Campbell, M.K., G. Piaggio, D.R. Elbourne, and D.G. Altman. 2012. "Consort 2010 Statement: Extension to Cluster Randomized Trials." *British Medical Journal* 345:e5661 doi: http://dx.doi.org/10.1136/bmj.e5661

Casey, Katherine, Rachel Glennerster, and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *The Quarterly Journal of Economics* 127 (4): 1755–1812.

Clark, Heather D., George A. Wells, Charlotte Huët, Finlay A. McAlister, L. Rachid Salmi, Dean Fergusson, and Andreas Laupacis. 1999. "Assessing the Quality of Randomized Trials: Reliability of the Jadad Scale." *Controlled Clinical Trials* 20 (5): 448–52.

Council of Economic Advisors. 2014. *Economic Report of the President*. US Government Printing Office.

Danzon, Patricia M., Sean Nicholson, and Nuno Sousa Pereira. 2005. "Productivity in Pharmaceutical–biotechnology R&D: The Role of Experience and Alliances." *Journal of Health Economics* 24 (2): 317–39.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48: 424–55.

DiNardo, John, and David S. Lee. 2011. "Program Evaluation and Research Designs." *Handbook of Labor Economics* 4: 463–536.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." *Handbook of Development Economics* 4: 3895–3962.

Dwan, Kerry, Douglas G. Altman, Juan A. Arnaiz, Jill Bloom, An-Wen Chan, Eugenia Cronin, Evelyne Decullier, Philippa J. Easterbrook, Erik Von Elm, and Carrol Gamble. 2008. "Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias." *PLoS One* 3 (8): e3081.

Ernst, Edzard, and Adrian R. White. 1998. "Acupuncture for Back Pain: A Meta-Analysis of Randomized Controlled Trials." *Archives of Internal Medicine* 158 (20): 2235.

Franco, Annie, Neil Malhotra, and Gabor Simonovits. 2014. “Publication Bias in the Social Sciences: Unlocking the File Drawer.” *Science* 345 (6203): 1502–5. doi:10.1126/science.1255484.

Frijters, Paul, Tao Sherry Kong, and Elaine M. Liu. 2015. *Who Is Coming to the Artefactual Field Experiment? Participation Bias among Chinese Rural Migrants*. Working Paper 20953. National Bureau of Economic Research. http://www.nber.org/papers/w20953.

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press. http://books.google.com/books?hl=en&lr=&id=7YyGAAAAQBAJ&oi=fnd&pg=PP1&dq=Running+Randomized+Evaluations:+A+Practical+Guide&ots=pOBGunnf3W&sig=8sIwBSkk_5AJC9KB4ZmoFm95W0g.

Gluud, Lise Lotte. 2006. “Bias in Clinical Intervention Research.” *American Journal of Epidemiology* 163 (6): 493–501.

Goldacre, Ben. 2014. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. Macmillan. https://books.google.com/books?hl=en&lr=&id=444XAwAAQBAJ&oi=fnd&pg=PP2&dq=bad+science+goldacre&ots=TSNYgfemlA&sig=jJMjle1SRLPAlRBum2f4b5_sFxE.

Guyatt, G. H., S. O. Pugsley, M. J. Sullivan, P. J. Thompson, L. Berman, N. L. Jones, E. L. Fallen, and D. W. Taylor. 1984. “Effect of Encouragement on Walking Test Performance.” *Thorax* 39 (11): 818–22.

Guyatt, G. H., E. J. Mills and Elbourne D. 2008. “In the Era of Systematic Reviews, Does the Size of an Individual Trial Still Matter?” PLoS Medicine 5 (1). doi: 10.1371/journal.pmed.0050004

Haahr, Mette Thorlund, and Asbjørn Hróbjartsson. 2006. “Who Is Blinded in Randomized Clinical Trials?” *The Cochrane Collaboration Methods Groups Newsletter* 3: 14.

Heckman, James J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica: Journal of the Econometric Society*, 153–61.

Higgins, Julian PT, Sally Green, and Cochrane Collaboration. 2008. *Cochrane Handbook for Systematic Reviews of Interventions*. Vol. 5. Wiley Online Library. http://onlinelibrary.wiley.com/doi/10.1002/9780470712184.fmatter/summary.

Hutton, J. L., and Paula R. Williamson. 2000. “Bias in Meta-Analysis due to Outcome Variable Selection within Studies.” *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 49 (3): 359–70.

Jüni, Peter, Douglas G. Altman, and Matthias Egger. 2001. “Assessing the Quality of Controlled Clinical Trials.” *BMJ* 323 (7303): 42–46. doi:10.1136/bmj.323.7303.42.

Jüni, Peter, Anne Witschi, Ralph Bloch, and Matthias Egger. 1999. “The Hazards of Scoring the Quality of Clinical Trials for Meta-Analysis.” *JAMA: The Journal of the American Medical Association* 282 (11): 1054–60.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. “Experimental Analysis of Neighborhood Effects.” *Econometrica* 75 (1): 83–119.

Kodrzycki, Yolanda K., and Pingkang Yu. 2006. “New Approaches to Ranking Economics Journals.” *Contributions in Economic Analysis & Policy* 5 (1). http://www.degruyter.com/view/j/bejeap.2005.5.issue-1/bejeap.2006.5.1.1520/bejeap.2006.5.1.1520.xml.

Leonard, Kenneth and Melkiory Masatu. 2006. “Outpatient Process Quality Evaluation and the Hawthorne Effect.” *Social Science and Medicine* 63 (9): 2330–2340.

Manski, Charles. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

McAmbridge, J., J. Witton and J. D.R. Elbourne. 2014. “Systematic Review of the Hawthorne Effective: New Concepts are Needed to Study Research Participation Effects.” *Journal of Clinical Epidemiology* 67 (3): 267-277. doi:10.1016/j.jclinepi.2013.08.015.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31. doi:10.1126/science.1245317.

Moher, D., K. F. Schulz, and D. G. Altman. 2001. "CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials." *Annals of Internal Medicine* 134: 657–62.

Moore, R. A., David Gavaghan, M. R. Tramer, S. L. Collins, and H. J. McQuay. 1998. "Size Is Everything–large Amounts of Information Are Needed to Overcome Random Effects in Estimating Direction and Magnitude of Treatment Effects." *Pain* 78 (3): 209–16.

Moore, R. A., M. R. Tramèr, D. Carroll, P. J. Wiffen, and H. J. McQuay. 1998. "Quantitative Systematic Review of Topically Applied Non-Steroidal Anti-Inflammatory Drugs." *British Medical Journal* 316 (7128): 333.

Noseworthy, John H., George C. Ebers, Margaret K. Vandervoort, R. E. Farquhar, Elizabeth Yetisir, and R. Roberts. 1994. "The Impact of Blinding on the Results of a Randomized, Placebo-Controlled Multiple Sclerosis Clinical Trial." *Neurology* 44 (1): 16–16.

Olken, Ben. 2015. "Promises and Perils of Pre-Analysis Plans." *Journal of Economic Perspectives* 29(3): 61-80. doi: 10.1257/jep.29.3.61

Oster, Emily. 2013. *Unobservable Selection and Coefficient Stability: Theory and Validation*. National Bureau of Economic Research. http://www.nber.org/papers/w19054.

Oxman, Andrew D., and Gordon H. Guyatt. 1992. "A Consumer's Guide to Subgroup Analyses." *Annals of Internal Medicine* 116 (1): 78–84.

Parker, Ian. 2010. "The Poverty Lab: Transforming Development Economics, One Experiment at a Time." *New Yorker* 17: 79–89.

Plint, Amy C., David Moher, Andra Morrison, Kenneth Schulz, Douglas G. Altman, Catherine Hill, and Isabelle Gaboury. 2006. "Does the CONSORT Checklist Improve the Quality of Reports of Randomised Controlled Trials? A Systematic Review." *Medical Journal of Australia* 185 (5): 263.

Rothwell, Peter M. 2006. "Factors That Can Affect the External Validity of Randomised Controlled Trials." *PLoS Hub for Clinical Trials* 1 (1): e9.

Schulz, Kenneth F., Douglas G. Altman, and David Moher. 2010. "CONSORT 2010 Statement: Updated Guidelines for Reporting Parallel Group Randomised Trials." *BMC Medicine* 8 (1): 18.

Schulz, Kenneth F., Iain Chalmers, Richard J. Hayes, and Douglas G. Altman. 1995. "Empirical Evidence of Bias." *JAMA: The Journal of the American Medical Association* 273 (5): 408–12.

Soares, Heloisa P., Stephanie Daniels, Ambuj Kumar, Mike Clarke, Charles Scott, Suzanne Swann, and Benjamin Djulbegovic. 2004. "Bad Reporting Does Not Mean Bad Methods for Randomised Trials: Observational Study of Randomised Controlled Trials Performed by the Radiation Therapy Oncology Group." *BMJ* 328: 22–25.

Spiegelhalter, David J., and Nicola G. Best. 2003. "Bayesian Approaches to Multiple Sources of Evidence and Uncertainty in Complex Cost-Effectiveness Modelling." *Statistics in Medicine* 22 (23): 3687–3709.

Temple, Robert, and Gordon W. Pledger. 1980. "The FDA's Critique of the Anturane Reinfarction Trial." *The New England Journal of Medicine* 303 (25): 1488.

The Cochrane Collaboration. 2010. "The Cochrane Collaboration, Home - The Cochrane Library." http://www.thecochranelibrary.com/view/0/index.html.

Thompson Reuters. 2010. "Thompson Reuters, ISI Web of Knowledge Journal Citation Reports for Medicine, General & Internal." http://admin-apps.isiknowledge.com/JCR/JCR.

Vader, J. P. 1998. "Randomised Controlled Trials: A User's Guide." *British Medical Journal* 317 (7167): 1258.

Vivalt, Eva. 2015. "How Much Can We Generalize from Impact Evaluations?" *Mimeo, New York University*.

Wood, L., M. Egger, L. L Gluud, K. F Schulz, P. Juni, D. G Altman, C. Gluud, R. M Martin, A. J.G Wood, and J. A.C Sterne. 2008. “Empirical Evidence of Bias in Treatment Effect Estimates in Controlled Trials with Different Interventions and Outcomes: Meta-Epidemiological Study.” *British Medical Journal*.

Wood, Lesley, Matthias Egger, Lise Lotte Gluud, Kenneth F. Schulz, Peter Jüni, Douglas G. Altman, Christian Gluud, Richard M. Martin, Anthony JG Wood, and Jonathan AC Sterne. 2008. “Empirical Evidence of Bias in Treatment Effect Estimates in Controlled Trials with Different Interventions and Outcomes: Meta-Epidemiological Study.” *British Medical Journal* 336 (7644): 601–5.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT press. https://books.google.com/books?hl=en&lr=&id=yov6AQAAQBAJ&oi=fnd&pg=PP1&dq=wooldridge&ots=iWfWDJGCUT&sig=i25YjbTCoCNHkO7QvosahQ7whuA.

Yusuf, Salim, Janet Wittes, Jeffrey Probstfield, and Herman A. Tyroler. 1991. “Analysis and Interpretation of Treatment Effects in Subgroups of Patients in Randomized Clinical Trials.” *JAMA: The Journal of the American Medical Association* 266 (1): 93–98.

Zwane, Alix Peterson, Jonathan Zinman, Eric Van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, et al. 2011. “Being Surveyed Can Change Later Behavior and Related Parameter Estimates.” *Proceedings of the National Academy of Sciences* 108 (5): 1821–26.

Figures and tables

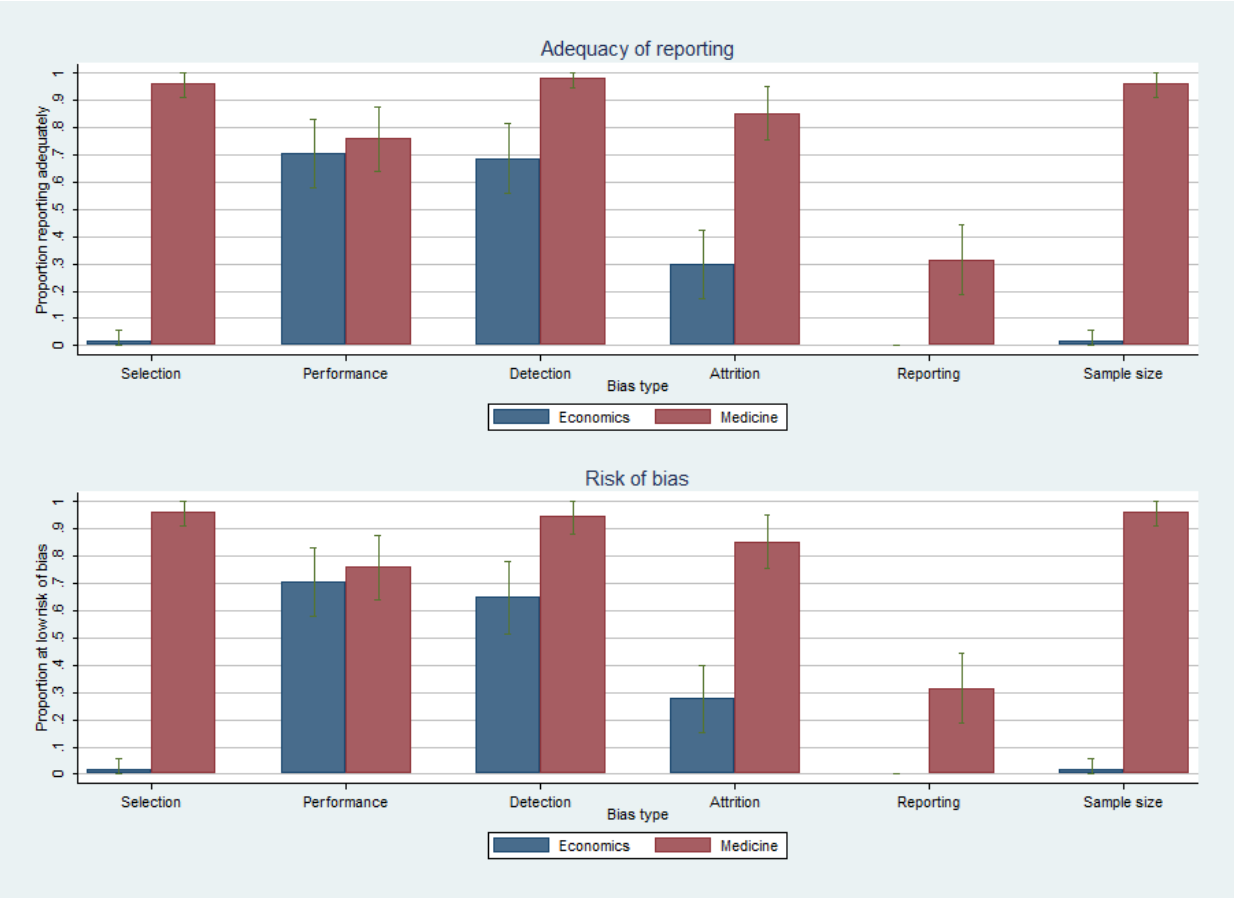

Figure 1. Assessment results overall, by field

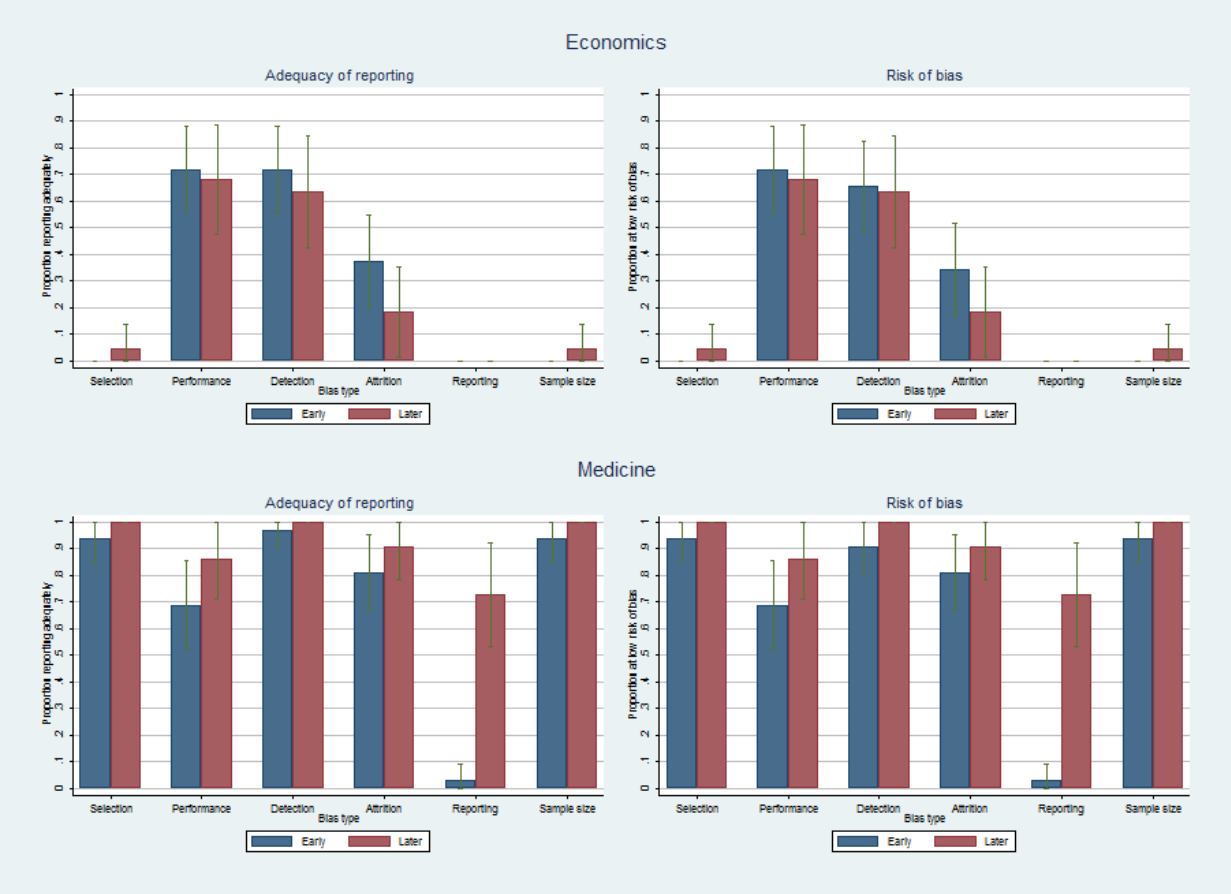

Figure 2. Assessment results separated by date of publication and field

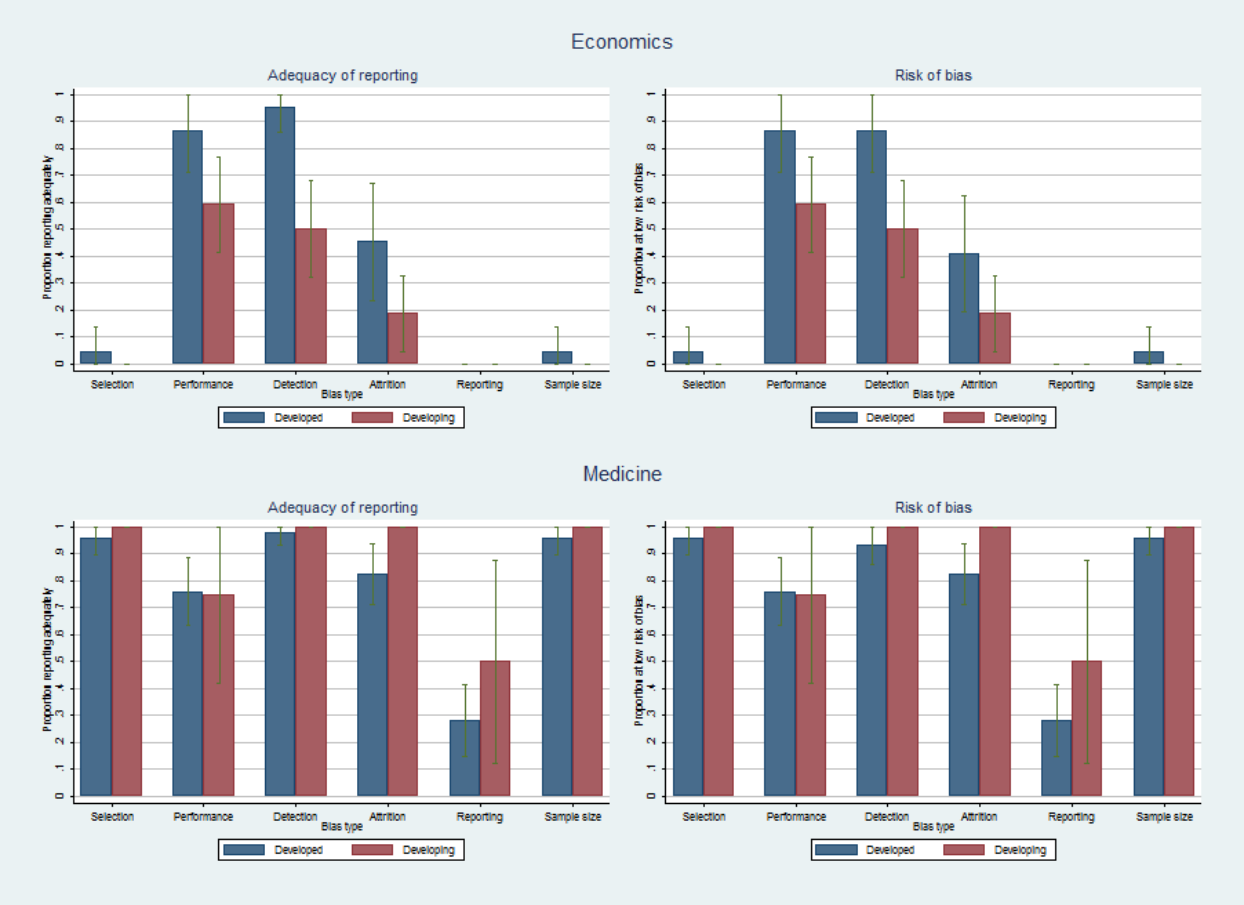

Figure 3. Assessment results from developing and developed countries, by field

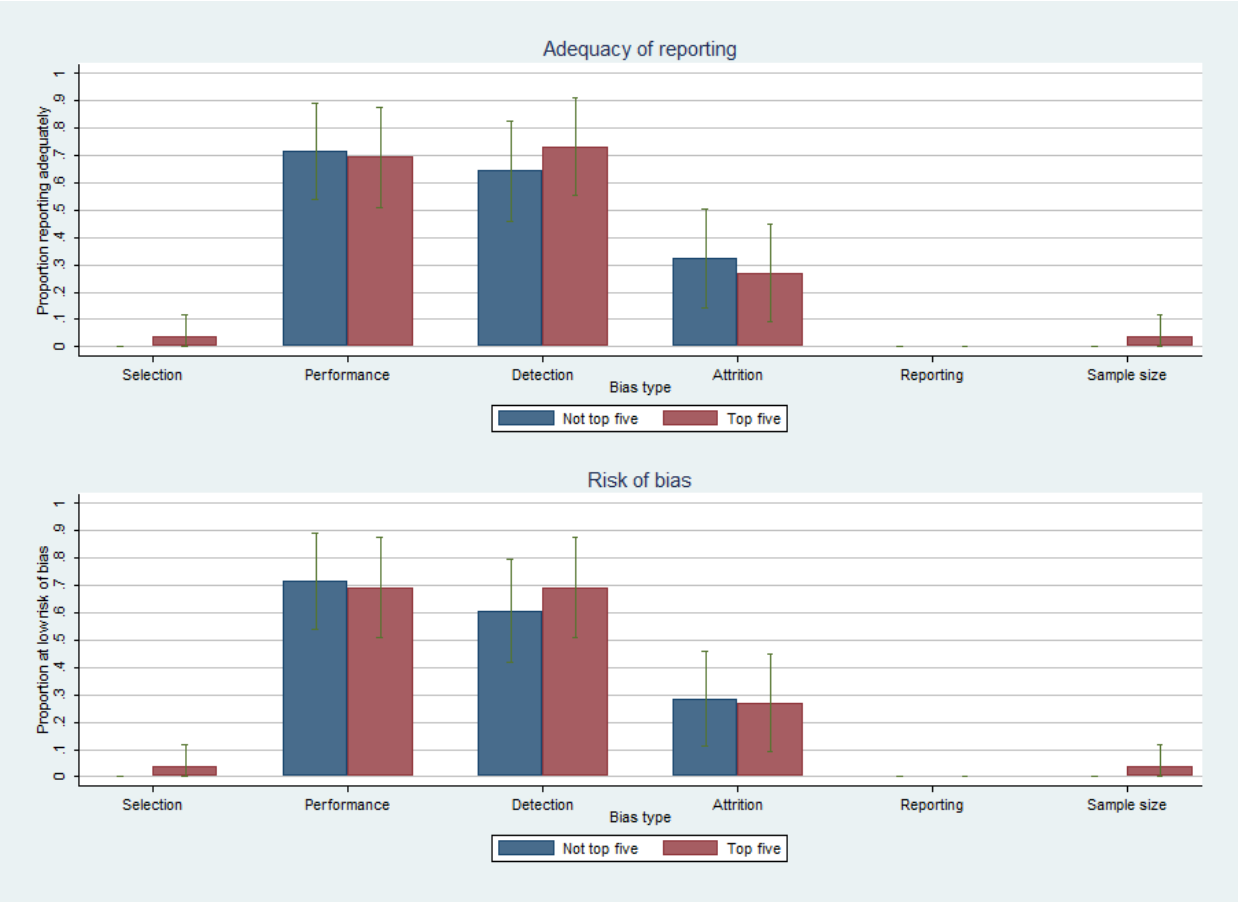

Figure 4. Assessment results by ranking of journal, economics only

Table 1—Assessment results by issue/bias and field

		Economics (n = 54)		**Medicine (n = 54)**		**P-value of Chi-square test**	
Bias	*Issue*	*Number reporting adequately*	*Number at low risk of bias*	*Number reporting adequately*	*Number at low risk of bias*	*Reporting*	*Risk of bias*
Selection	Randomization	34	29	52	52	<.001	<.001
Selection	Flow of participants	20	20	40	40	<.001	<.001
Selection	Baseline demographics	48	46	53	52	.051	.046
Attrition	Flow of participants	17	17	51	51	<.001	<.001
Attrition	Intent-to-treat	35	33	47	47	.007	.002
Performance	Data collection	43	43	51	51	.022	.022
Performance	Participant behavior	45	44	42	42	.466	.633
Detection	-	37	35	53	51	<.001	<.001
Reporting	Protocol/analysis plan	0	0	50	50	<.001	<.001
Reporting	Outcomes	0	0	49	49	<.001	<.001
Reporting	Interpretation of results	12	12	19	19	<.001	<.001
Sample size	-	34	34	51	51	<.001	<.001
Aggregated to bias level							
Selection		9	12	39	40	<.001	<.001
Attrition		15	16	46	46	<.001	<.001
Performance		38	38	41	41	.515	.515
Detection		37	35	53	51	<.001	<.001
Reporting		0	0	17	17	<.001	<.001
Sample size		1	1	52	52	<.001	<.001