ISSN 2042-2695

CENTRE *for* ECONOMIC PERFORMANCE

CEP Discussion Paper No 1240

September 2013

Risk and Evidence of Bias in Randomized Controlled Trials in Economics

Alex Eble, Peter Boone and Diana Elbourne

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE ■

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Abstract

The randomized controlled trial (RCT) has been a heavily utilized research tool in medicine for over 60 years. Since the early 2000's, large-scale RCTs have been used in increasingly large numbers in the social sciences to evaluate questions of both policy and theory. The early economics literature on RCTs invokes the medical literature, but seems to ignore a large body of this literature which studies the past mistakes of medical trialists and links poor trial design, conduct and reporting to exaggerated estimates of treatment effects. Using a few consensus documents on these issues from the medical literature, we design a tool to evaluate adequacy of reporting and risk of bias in RCT reports. We then use this tool to evaluate 54 reports of RCTs published in a set of 52 major economics journals between 2001 and 2011 alongside a sample of reports of 54 RCTs published in medical journals over the same time period. We find that economics RCTs fall far short of the recommendations for reporting and conduct put forth in the medical literature, while medical trials stick fairly close to them, suggesting risk of exaggerated treatment effects in the economics literature.

Keywords: randomized controlled trials, field experiments, bias, treatment effect estimates
JEL Classifications: C9, C90, C93, C10, C12, C18

This paper was produced as part of the Centre's Wellbeing Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

Acknowledgements

Alex Eble (alexander.eble@brown.edu) is a PhD student in economics at Brown University. Prior to his PhD studies he served as trial coordinator for two large cluster-randomised trials in rural India evaluating primary health care and education interventions. He has also lived, worked and conducted research in China and the UK focusing on education and well-being.

Peter Boone (pb@effint.org) is a research associate at the Centre for Economic Performance, London School of Economics, and a non-resident Senior Fellow at the Peterson Institute for International Economics, Washington D.C. He is also chairman of Effective Intervention, a UK based charity which has sponsored and implemented numerous randomized controlled trials examining methods to reduce child mortality and improve children's literacy and numeracy in poor regions of India and Africa.

Diana Elbourne (Diana.Elbourne@LSHTM.ac.uk) is Professor of Healthcare Evaluation in the Medical Statistics Department at the London School of Hygiene and Tropical Medicine. She has a background in social sciences and statistics, and has been conducting both methodological and applied research in randomised controlled trials (RCTs) for over 30 years. She has been a key member of the Consolidated Standards of Reporting Trials (CONSORT) group (http://www.consort-statement.org) since 1999, and is a co-author of both the 2001 and 2010 revisions to the CONSORT Statement and a lead author to its extension to two RCT designs (Cluster trials, and Non-inferiority and Equivalence trials, published in 2012)

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

Section I: Introduction

Assignment of treatment to different groups and subsequent comparison of outcomes dates as far back in history as the Old Testament, in which King Nebuchadnezzar is said to have ordered a group of his subjects to eat rich meat and drink wine while another group was made to adhere to vegetarianism in order to evaluate the merits of the two diets. (1 Daniel 11-16, New International Version) Though many approximations of the randomized controlled trial (RCT) have been conducted since, (Twyman 2004) the dawn of the current era of the RCT is a set of articles published in Journal of the American Medical Association in the 1940s. (Bell 1948) Hundreds of thousands of trials have been conducted since. The method has been shown by several studies to yield less biased treatment effect estimates than observational studies and as a result has been adopted in several scientific fields as the "gold standard" of evidence. (Vader 1998)

Despite this acclaim, decades of use and scrutiny have revealed numerous potential problems in the execution and review of RCTs centering on a set of six potential biases related to trial conduct and analysis. The problems central to each of the six concerns have been associated with exaggerated treatment effects relative to studies whose design anticipates and attempts to prevent such problems. Stemming from these findings, a few consensus documents have been developed to provide guidance on how best to design, conduct and report trials in order to minimize the risk of such problems biasing the results.

In the past decade, the RCT has been widely adopted by economists - largely on the virtue of its "clean" identification of causal relationships - and has been used by economists to evaluate hundreds of questions of both academic and policy interest. (Parker 2010) Though economists mention and often cite the medical literature as the inspiration for this approach, (Abhijit Banerjee 2007) surprisingly few published reports of economics trials published before 2010 reference any of the wealth of medical articles on the pitfalls which have been shown to lead to biased results or any of the articles on the means by which to reduce such biases.

Our research question is this: have trialists in economics taken the necessary steps to avoid the bias-inducing pitfalls that the medical literature has identified? Below, we briefly summarize the medical literature on bias in RCTs. We have used this literature to develop an instrument (henceforth, the "grid") with which to evaluate adequacy of reporting and risk of the six aforementioned biases (henceforth referred to simply as "bias") in published accounts of RCTs. Though we recognize it is an open question to what extent the standards from medicine can be meaningfully applied to economics, we argue that the medical standards offer a very clear link between certain RCT design and conduct decisions and treatment effect exaggeration, and that there is no reason not to use this information. After the discussion, we then use the grid to evaluate a set of journal articles documenting the results of RCTs in both economics and medicine.[1] We find that many of the economics articles provide insufficient information for the reader to assess the quality of the evidence presented and several others fall into the same traps that have previously skewed the results of trials in medicine. We finish by suggesting a similar set of guidelines for trialists in economics to follow when conducting and evaluating RCTs and offering a few paths for future research.

[1] Economists have long been concerned with the same issues of identifying causality that led medical scientists to use RCTs, and there is a rich history of economists conducting experiments, both in the laboratory and beyond. Our analysis focuses exclusively on the use of prospectively designed, relatively large-scale RCTs in economics which have been in vogue only for the last decade and whose mission, arguably, mirrors the "Phase III" trial in medicine. They differ from previous experiments in economics in both their scale and objectives.

Section II: Trials in Medicine and Economics

The history of randomized trials is well documented elsewhere, (Meldrum 2000; Collier 2009) so we provide only a brief discussion of their development to motivate our analysis. Though the first parallel group study ostensibly dates to pre-Christian times as discussed in Section I, trials have only been broadly accepted by the medical community since the 1940s. As early as 1980 the RCT was recognized for its superior identification of causal relationships relative to other research designs, (Vader 1998) confirmed empirically in a series of meta-analyses which showed that nonrandomized studies yielded larger effect sizes than those found in randomized trials.(Ioannidis et al. 2001)

Subsequent analysis of the evidence provided by RCTs revealed that errors in design or analysis could lead to exaggerated treatment effect estimates in trials. A series of studies investigated the relationship between methodological quality of RCTs and measured effect size, beginning with a landmark 1995 article which found that trials with inadequately concealed treatment allocation estimated up to 30% larger treatment effects than well designed studies with adequate allocation concealment. (Schulz et al. 1995) This finding and others similar to it instigated a larger movement to improve and standardize both methods of reporting RCTs and methods of scrutinizing them.

In the 1990s, two groups began independently working on establishing a set of reporting standards to be used in publication of randomized trials, the goal of which was to ensure that readers of articles reporting the results of RCTs had sufficient information to confirm or refute that the trial had in fact been carried out in a manner which would yield unbiased results. Their combined efforts resulted in the CONSORT Statement, 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 Cochrane Collaboration, another arm of this movement, is an international organization which 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 about how to conduct these reviews which includes a section on how to evaluate the quality of evidence provided by RCTs. The handbook, which is updated frequently, has been used in 6,200 systematic reviews of trials, which have together assessed the quality of evidence in hundreds of thousands of scholarly articles. (The Cochrane Collaboration 2010)

The Cochrane handbook and CONSORT Statement offer a thorough discussion of the six problems 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) The remit of each of these issues is vast and thorough exploration of any of them is beyond the scope of this study. Instead, we discuss each issue briefly and cite a few major studies which demonstrate the implications of study design which fails to address the potential pitfalls associated with it.

Selection bias

Selection bias in the context of trials is the concern that systematic differences exist between treatment groups at the outset of the trial, usually due to individuals either tampering with or predicting the allocation sequence. A review of several meta-analyses which aggregated the results of Schulz and others found that "odds ratios were approximately 12 percent more positive in trials without adequate allocation sequence generation" and that "trials without adequate allocation concealment were approximately 21 percent more positive than trials with adequate allocation concealment". (Gluud 2006)

The CONSORT Statement asserts that “authors should provide sufficient information that the reader can assess the methods used to generate the random allocation sequence and the likelihood of bias in group assignment.” The Cochrane Handbook states

> “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.”

In any proposed randomization sequence, there is risk that it can be either tampered with by someone involved in allocation (e.g. covertly breaking the sequence in order to assign the intervention to those seen as more needy) or simply inadvertently deterministic due to poor design (e.g. either by assigning treatment using a sequence that could be predicted by participants who would then selectively enroll or by deterministically assigning participants to groups by a rule relying on a nonrandom characteristic such as birth date), both of which can result in nonrandom treatment allocation and therefore biased treatment effect estimates. (Wood et al. 2008; Schulz, Altman, and Moher 2010)

In addition, in this study we add to the traditional notion of selection bias the concern that systematic differences arise between the stated population from which the sample is drawn and the randomized participants . Though this is traditionally considered the realm of generalizability, the potential problem here is that any such difference, if not fully disclosed, could result in biased treatment effect estimates for the specified population of interest.

If, for example, we are told that the population is from a specific sample (e.g. all smokers who smoke two or more cigarettes per day) but the final population sampled from differs substantially from the specified population (e.g. only smokers who smoke between 2 and 5 cigarettes per day), then the treatment effect (e.g. of the efficacy of a low-intensity stop-smoking intervention) we observe may differ from the actual treatment effect of the intervention for the specified population. Such a difference in reported and actual treatment effect estimates from a trial in such circumstances is functionally similar to a treatment effect bias arising from the other problems discussed in this section. It is therefore imperative that trialists specify exactly who is screened for eligibility, who is eligible, who is enrolled in the trial and who is excluded in order to prevent such a discrepancy in reported and actual population-specific treatment effects.

Performance bias

Also known as the set of “Hawthorne” and “John Henry” effects, or concomitant treatment bias, performance bias is the tendency for participants to change their behavior or responses to questions because they are aware of being in a trial and of the treatment allocation. In many medical trials blinding of participants is used to minimize this type of bias, as a participant unaware of allocation status is unable to act upon that knowledge. Discovery of allocation status in trials which were intended to be blinded has been linked to skewed results, a famous example of which is a 1975 study of the effects of Ascorbic Acid on the common cold, whose main result was that the (small) measured treatment effect was likely due to such bias. (Karlowski et al. 1975) The Cochrane Handbook states:

> “Lack of blinding of participants or healthcare providers could bias the results by affecting the *actual* outcomes of the participants in the trial. This may be due to a lack of expectations in a control group, or due to differential

> behaviours across intervention groups...lack of blinding might also lead to bias caused by additional investigations or co-interventions regardless of the type of outcomes, if these occur differentially across intervention groups."

In the cases where blinding is impossible, it is essential to recognize and address concerns of bias resulting from knowledge of treatment allocation, as knowledge of differential treatment status is likely to be linked to differential subsequent outcome-related actions (e.g. seeking or providing alternative treatments).

The likely direction of performance bias is ambiguous. On the one hand, the placebo effect is well known. A meta-analysis of studies of acupuncture treatment on back pain which showed that while acupuncture was superior to control interventions, (unblinded studies) it could not be proven to be superior to sham-interventions (blinded studies). (Ernst and White 1998) Conversely, in an RCT evaluating a medical intervention, if participants in the control group were aware of the intervention group treatment strategy, we might expect them to be more likely to seek outside care than before as a result of said awareness, which would introduce a systematic downward bias on treatment effect estimates.

Risk of such bias is difficult to control for in many trials, particularly those in economics, as blinding is often impossible and the counterfactual - "what would the group have done if they had not been aware of their treatment allocation?" - cannot be answered. Nonetheless, the CONSORT Statement maintains that the possibility that knowledge of treatment allocation could skew behavior of the two groups differentially should be explicitly addressed in reports of RCTs in order to accurately assess the quality of data the trial provide.

Detection bias

As with performance bias, detection bias (or assessment bias, as it is sometimes referred to in the medical literature) is concerned primarily with blinding. In this case, however, the concern is about those collecting the data, not those providing it. The Cochrane Handbook warns that if "outcome assessors are aware of assignments, bias could be introduced into *assessments* of outcome, depending on who measures the outcomes." (Higgins, Green, and Collaboration 2008) (italics original) A trial evaluating the impact of blinding data assessors on measured treatment effect showed that preconceptions of treatment efficacy and placebo effects can have similar effects on data collectors and assessors as they do on participants. (Noseworthy et al. 1994) The CONSORT Statement adds that "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." (D. Moher, Schulz, and Altman 2001) Evidence of detection bias has also been found 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%. (Guyatt et al. 1984) Unblinded trials which are not scrupulous in hiring third-party data collectors and training them to avoid these problems (as well as reporting these efforts) are therefore at higher risk of detection bias. Though all outcome assessments can be influenced by lack of blinding, there are particular risks of bias with more subjective outcomes (e.g. severity of pain or satisfaction with care received). It is therefore recommended in these instruments to consider how subjective an outcome is when considering blinding. Lack of blinding has been associated with a 30% exaggeration in treatment effect estimates in a meta-analysis of studies with subjective outcomes. (Wood et al. 2008)

Attrition bias

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. One way of perceiving this concern is as an extension of the concerns outlined in the selection bias section taken forward to the execution and completion of the trial. Loss of participants can come from any number of reasons: drop-out, missing data, refusal to respond, death, or any exclusion rules applied after randomization. As explained in an article discussing this bias: "any analysis which omits patients is open to bias because it no longer compares the groups as randomised [sic]." (Lewis and Machin 1993)

One particularly salient example of post-hoc exclusion creating bias is the Anturane Trials, wherein the authors excluded those participants who died during the course of the trial, despite the fact that mortality rates differed highly between control and intervention groups. The initial article from these trials showed a significant effect of the drug, but subsequent analyses which included participants according to randomization status failed to reject the null of no treatment effect. (Temple and Pledger 1980)

Reporting bias

Perhaps the most insidious of the problems facing those reading the reports of RCTs, reporting bias is the concern that authors present only a subset of outcomes or analyses and, as a result, the reader is left with an incomplete and often skewed understanding of the results. The more serious risk is that this bias will lead to many false positive conclusions about the efficacy of treatment and this, in turn, will lead to misinformed care or policy. The likelihood of this risk has been identified in a review of oncology articles published in two major medical journals, (Tannock 1996) and a more recent article confirmed this finding in three separate meta-analyses, finding 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 recent meta-analysis of studies on anthelminth therapy and treatment for incontinence found additional evidence that "more outcomes had been measured than were reported", and 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 take two major steps. One, they require that a trial and brief protocol be registered with a central, third-party database before the study begins. The protocol documents the plan for conduct of the trial, the intended sample size, the outcomes and the analyses that the trialists will undertake at the end. This ensures continuity in the conduct of the trial, as any post-hoc changes that are made, potentially in favor of presenting more interesting results, would contradict the publicly available plan for action. Upon consideration for publication, journal editors and peer reviewers can use the protocol to check for this.

The second is to create a statistical analysis plan (often called a "pre-analysis plan" in economics (Casey, Glennerster, and Miguel 2012)) which specifies before the beginning of the trial which analysis will be defined as the primary endpoint or primary analysis. The construction of t-test or similar comparison of means with a 95% confidence interval is such that conducting 20 such analyses will on average yield one "significant" result by virtue of chance alone. To prevent authors from running analyses *ad infinitum* and publishing only those which are significant, both the protocol and subsequent report of the article must report which analysis is primary and thus given the highest credence.

In this process, additional labels of "secondary" (pre-planned, but not the primary analysis) and "exploratory" (conceived of after the data was collected and examined) outcomes are required to be assigned to the remaining presented results. This allows the reader to differentiate between analyses that the authors planned before the study and analyses which were conceived after the authors were able to examine the data. Exploratory analyses are still seen as informative, but are given less weight than pre-specified analyses, as there is a high risk of false-positive results in *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) While there are tools available which can help mitigate some types of the multiple comparison problem, (Kling, Liebman, and Katz 2007) a recent study from the economics literature documents how separate and contradictory erroneous conclusions could have been drawn from a randomized experiment in Sierra Leone in the absence of pre-specification of endpoints. (Casey, Glennerster, and Miguel 2012)

Sample size bias

The first concern here is that an insufficiently large sample size can lead to imprecise estimation and therefore to misleading conclusions. The CONSORT Statement describes one 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."

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) Two other studies found that small sample sizes were likely to overstate the effect size because of the heightened influence of outliers in these cases. (Moore, Gavaghan, et al. 1998; Moore, Tramèr, et al. 1998) To guard against these problems, both the CONSORT Statement and Cochrane Handbook expect trialists to conduct sample size calculations before collecting any data and report these calculations in trial publications.

Scrutiny of these issues

As mentioned earlier, adherence to the CONSORT Statement guidelines is now required by many journal editors for publication. (Schulz, Altman, and Moher 2010) Articles which are successfully published in peer reviewed journals are again scrutinized by Cochrane Collaboration contributors during the conduct of systematic reviews. This repeated scrutiny has resulted in a reduction, over time, in the presence of the biases described above in medical RCT reports. (Plint et al. 2006) In line with this finding, the FDA uses a similar set of standards to approve the sale of pharmaceuticals for public sale and consumption. For a drug to be approved by the FDA, it must pass three "phases" of trial with increasing scrutiny at each phase (i.e. phase II trials have a higher burden of proof than phase I but less than phase III). Looking at the progress of different pharmaceuticals through this process, it is clear that these standards have substantial impact on the results of a given study: of the trials that enter phase II, less than 50% pass the two phase III trials usually necessary for FDA approval. (Danzon, Nicholson, and Pereira 2005)

Economics trials and our motivation

As discussed in the introduction, academics in pure (as opposed to medical) economics departments have witnessed a surge in the use and popularity of large scale RCTs in the last ten years. (Parker 2010) Review of the bibliographies of articles in economics journals reporting the results of these RCTs, however reveals that many of the trials conducted to date have not explicitly drawn from the health literature on how to minimize bias in such experiments in the ways discussed above. As a result, we are concerned that economics trials unnecessarily risk stumbling into the same pitfalls which have plagued medical trials for the past sixty years.

In the section that follows, we describe the development and application of the grid, an instrument which uses the insights from the literature cited above as its main source. We are eager to acknowledge that the goals of economics trials are not identical to those of medical trials and that it is an important question to ask how the metrics used to evaluate them should also differ. In light of this concern, the grid does not perfectly mirror the CONSORT Statement or Cochrane Handbook. Rather, it incorporates those suggestions which seem most appropriate to economics and excludes others which are either inappropriate for most economics trials (e.g. strict views on blinding) or insufficiently objective (e.g. issues surrounding generalizability).

As for the criteria which remain, we contend that there are two justifications for applying them to the economics literature. One is that we see this as a $100 bill lying on the ground. The medical literature has carefully identified a set of well-defined concerns and shown that lack of attention to them yields bias in treatment effect estimates. There seems little reason not to draw on this experience. We also recognize that evidence from many recent economic RCTs has been used to inform economic and social policy in both developing and developed countries. As these policies affect a large proportion of lives globally, we argue that standards of equal rigor should be applied to these policy decisions as are applied to the decision whether to approve a wide array of interventions in the health arena.

Section III: Methodology

In this paper, we hope to answer the following research question: are the recent reports of RCTs in economics providing readers with sufficient information to assess the quality of evidence provided by the experiment (henceforth: are they adequately reporting how the trials were conducted) and is there evidence that authors take the necessary steps to minimize the risk of the biases that medical trialists have encountered? To answer this question, we developed a reporting and bias evaluation tool using a subset of the standards and guidelines in the CONSORT Statement and the Cochrane Handbook. We then collected all economics articles reporting on trials which mention randomization in the title or abstract published in a set of 52 major peer reviewed journals between 2001 and 2011. To evaluate the validity of our grid and to provide a benchmark for our ratings of articles in economics, we randomly selected an equal number of articles from peer reviewed journals in medicine. Finally, we applied our grid to both sets of articles. Below we describe our grid, our article selection process, and the assessment process itself.

The grid

To systematise the assessment of articles, we developed a grid which addresses each of the issues discussed in section II, provides leading questions to assist the assessor in assessment, and facilitates data collection. The full grid is given in Appendix 1. It is designed to facilitate and collect assessments of adequacy of reporting and risk of bias in terms of the six biases. There are 13 broad “issues” spread across the six biases, and many of these contain several smaller questions. The task of the assessor is to answer each question by putting either a “√”

for yes or an “X” for no to the left of the question and, if at all possible, provide a page number or explanation in the comment and quote boxes to the right of the question to justify the assessment. The assessor then aggregates the assessments from questions to issues, and then aggregates from issues to an overall assessment for each of the six biases, separately for adequacy of reporting and risk of bias using a simple rule: if the article fails on any issue in terms of adequacy of reporting, then it fails for the overall adequacy of reporting of that bias (and similarly for the assessment of low risk of bias). The motivation for this structure is that each type of bias is complex, comprising several different concerns, each of which must be addressed to minimize the risk of a given bias. The result of this grading process was an assessment for each of the 13 issues and each of the 6 biases, whether the issue/bias was reported adequately, and whether or not there was low risk of bias associated with that issue/bias.

The studies

For this analysis, we collected a set of articles published in peer-reviewed journals in economics reporting the results of economics trials. The selection process was as follows:

1) Using the EconLit database, we searched for journal articles published between 2000 and 2009 that contained either the word randomized or randomization (or their alternative British spellings) in the title or abstract. A search conducted on July 6th, 2010 generated 527 results. This was amended on September 5th, 2012, with the results from a search which expanded the range of the original search to include papers from 2010 and 2011, which yielded 235 additional results.[2]
2) From these results, we further limited eligibility with two criteria:
    a. The first eligibility criterion was that an article had to report the results of a prospectively randomized study. This condition was incorporated in light of the fact that we are evaluating study design and so it would be inappropriate to include studies not specifically designed as trials (e.g. public lotteries or other 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) When the search was expanded in 2012, we also included studies published in the *American Economic Journal: Applied Economics* and the *American Economic Journal: Economic Policy* from their inception in 2009 to the end of 2011. This decision was made in light of their prestige and the volume of RCT reports published in them.

In total, this yielded 54 articles published between 2001 and 2011. A full list is provided in Appendix 2.

We randomly selected an equal number of articles reporting phase III trials published in three of the top peer-reviewed medical journals for grading.[3] This served two purposes –

[2] We recognize that this is not the universe of published RCTs but believe it is a good approximation. Scanning the table of contents of all the journals over the period would have been prohibitively time consuming and including the word “experiment” in the search terms raises the number of initial results well into the four digit range.

[3] We chose phase III trials as they are the most akin to the large-scale RCTs in economics which we are examining and are subject to the highest burden of proof. As described in the previous section, a medical intervention, pharmaceutical or otherwise, must pass two phase III trials to be approved by the FDA for public sale and consumption.

one, to ensure that our grading instrument 'worked'[4], and two, to provide a benchmark for how the "gold standard" in medicine would fare according to our standards. We drew our sample such that in each year with at least one eligible article in economics, there were selected an equal number of articles in medicine as there were eligible articles in economics. We chose to draw this sample of articles in medicine from the top three medical journals as classified by the Thompson *Journal and Citation Reports*' impact factor in general and internal medicine as of July 6th, 2010. (Thompson Reuters 2010) These journals are *The Lancet, The Journal of the American Medical Association,* and *The New England Journal of Medicine*. The decision to only consider articles from these three journals was made with two motives: one, for ease of processing, as there are thousands of RCT reports published each year and restricting the journals to these three still left us with approximately 350 each year and, two, in order to see how our grid fared evaluating the "gold standard" in medicine.

To obtain the medical RCT article sample, we used the following process:

1) We searched Pubmed (a database similar to Econlit indexing medical journals and their articles) for all articles reporting clinical trials in these three journals in years when there was also an eligible economics article published (all years in our range save 2002).
2) From this list, we then randomly selected a number of articles in a given year equal to the number of eligible articles in economics in that year. Randomization was performed by ordering the journal articles as they appeared in the search, assigning each article a random number between 0 and 1 using a random number generator, and then sorting the articles in ascending order by the magnitude of the randomly assigned number, selecting the first x articles required to achieve balance between the two fields.
3) We excluded Phase I and II trials in medicine as their methods, goals and sample size considerations are significantly different from Phase III trials, which, similar to the economics trials we are concerned with, are more often used to inform policy.

The final list of both sets of papers is given in Appendix 2. In both medicine and economics, if a trial generated more than one eligible publication, the article published earliest was selected. Other associated articles were only used to provide additional information for evaluation of the main article.

The assessment process

The grid was first piloted by all three authors and Miranda Mugford. Once the grid was finalized, two authors (AE/PB) first read each article and assessed the adequacy of reporting and risk of bias using the grid 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, presented in section IV. This method of individual grading followed by deliberation was adopted following the example of several meta-analyses in the medical literature, which find that while independent grading potentially provides better internal validity of the grid, the rate of agreement between graders in such processes is often low. (Clark et al. 1999) In practice, our mean rate of agreement on sub-issue assessment is greater than 85 percent.

In the analysis on risk of bias that follows, we group inadequacy of reporting (and therefore unclear risk of bias) with high risk of bias. While this is not ideal, unclear risk of

[4] 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 fail most medical trials on many biases (pass all of them on all issues), we would be concerned that the instrument was too strict (lenient).

bias sheds similar, if not as severe, doubts on the conclusions of the study in question. We draw this method from the landmark meta-analysis assessing study quality in medicine. (Schulz et al. 1995) We do not aggregate the individual scores to create an overall study-level score, as each section represents a separate concern, again following the lead of meta-analyses in medicine. (Spiegelhalter and Best 2003) As the issues in our analysis are diverse, bias-specific treatment effect estimate exaggeration magnitudes are likely to differ across biases.

Section IV: Analysis

In this section we compare our assessments of published articles in economics and medicine, in terms of adequacy of reporting and risk of bias. We find that the economics literature reports on the majority of these risks irregularly - for four of the six biases, less than 30 percent of the articles collected report adequately, and for no type of bias do more than three quarters of the articles report adequately. The pattern is largely similar for our assessments of risk of bias in economics articles. Though the relationship between adequacy of reporting and risk of bias is often direct, even among the subset of articles in which reporting is adequate there are many cases in which there is high risk of bias. For two of the six biases, all but two of the articles in economics that we include fail to report adequately and cannot be assessed as having low risk of bias. The medical literature, as expected, does much better, though for no bias do 100 percent of the articles report adequately or have low risk of bias.

Below, we show summary statistics of our assessments and then provide selected examples of concerns from the economics articles. Simple bar charts documenting performance of economics articles and medical articles in terms of adequacy of reporting and risk of bias are given in Appendix 3.1. Similar charts breaking down the assessments of each of the six biases by issue are given in Appendix 3.2. Table 1 below gives the data from Appendix 3.1 numerically alongside a two-tailed student's t test with heteroskedastic errors.

Table 1 – Performance of articles by issue and discipline

Bias	Issue	Economics articles passing N=54	Medical articles passing N=54	Chi-squared test p-value
Selection	Reporting	22.2%	74.1%	0.000
Selection	Risk of bias	16.7%	72.2%	0.000
Performance	Reporting	70.4%	75.9%	0.515
Performance	Risk of bias	70.4%	75.9%	0.515
Detection	Reporting	68.5%	98.2%	0.000
Detection	Risk of bias	64.8%	94.4%	0.000
Attrition	Reporting	29.6%	85.2%	0.000
Attrition	Risk of bias	27.8%	85.2%	0.000
Reporting[5]	Reporting	0.0%	81.5%	0.000
Reporting	Risk of bias	0.0%	81.5%	0.000
Imprecision	Reporting	1.9%	96.3%	0.000
Imprecision	Risk of bias	1.9%	96.3%	0.000

[5] Our initial instrument included a requirement for presenting an online table of "ancillary analyses" as one of the sub-issues in reporting bias. After the first round of grading, review of the literature and discussion with authors responsible for the CONSORT Statement, it was clear that this was a bad criterion, as requiring this unnecessarily penalized papers, both in economics and medicine, which performed no ancillary analyses and therefore had nothing to report. We do not use this sub-issue in our assessments of reporting bias, but leave it in our grid in the appendix for full disclosure.

Selection bias

Only 12 of the 54 eligible economics articles (22%) passed the reporting criteria for selection bias. For reference, 40 of the 54 eligible medical articles did so. The vast majority of papers in economics provided insufficient details on the process used to randomize, an ambiguity which leaves doubt as to whether the randomization processes used could have been deterministic or that an administrator or investigator could have corrupted the sequence. Five articles did report their means of randomization, but used clearly deterministic methods (for example, an alphabetic algorithm in one case and sorting by date of employment commencement in another) to assign treatment. Lack of information about the flow of potential participants in the trial was another major flaw in articles in economics. In the majority of the eligible articles published in economics journals, the numbers of participants screened for eligibility and excluded before and after eligibility was assessed were not given.

Performance bias

Sixteen of the 54 economics papers reported inadequately in terms of performance bias and an equal amount had high risk of bias. In most medical trials, this problem is often avoided by blinding participants to which treatment group they have been assigned to. In some instances, this is impossible, but when blinding is not feasible, the medical literature (and our grid) requires that the authors of the study discuss the potential for such bias and demonstrate that it was not in fact a risk. The economics papers which failed on these criteria almost uniformly neglected to address this concern and due to the design of their trial (e.g. use of subjective / self-reported endpoints) seemed at particular risk for the issue. It is important to note that we did not fail papers for not blinding – rather, a paper did not pass on adequacy of reporting if there was apparent risk of performance bias (e.g. alternative care-seeking as a result of knowledge of treatment status) which was not discussed. In an article which evaluated a program which gave cash transfers conditional on school enrolment, for example, there is a clear concern that participants assigned to the control group might change their behavior (by waiting to send children to school, for example, until the program was rolled out to all households) in light of their knowledge of their and others' treatment status. There was no mention of this concern in the article in question.

Detection bias

Shortcomings in terms of detection bias had to do with the identity of data collectors and the nature of data. Seventeen of the 54 economics articles failed on reporting and 19 on risk of this bias. Many of these trials collected data with individuals who may have had incentive to skew the data in favor of the intervention. Two articles explicitly mentioned using data collectors who were employed by the same company which administered the intervention. Several others neglected to say who collected the data, leaving doubt as to whether a similar conflict of interest could have biased the results.

Attrition bias

There are two interlinked concerns here – one is that participants dropped out during the course of the trial in a way that would destroy the balance between treatment groups achieved by randomization. The other concern is that the analyses run do not follow the "intent to treat" principle, which stipulates that all randomized participants be included in the final analysis. Only 17 of the 54 economics articles passed this criterion. More than 20 did not discuss exclusion of participants in the final analysis and almost all of these had widely varying numbers of observations in different versions of the same analysis, suggesting that selective exclusion of observations did in fact take place. Less than half of the articles we collected mentioned the intent to treat principle by name and, among those that did, several neglected to follow it. Many of these articles excluded groups of participants because they did not follow the protocol, and one paper threw out the second of two years of data collected because of contamination. While these concerns do not definitively show bias, they leave open the possibility for bias from attrition, an ambiguity that has been associated with exaggerated results in medical trials.

Reporting bias

No economics paper was adequately reported in terms of reporting bias, and therefore none could be assessed as having low risk of bias in this category. Our assessment attests to two phenomena. The first and foremost is the lack of both pre-specification of endpoints and registering a study and a brief protocol prior to implementation of the trial. As described in Section II, pre-specifying a primary endpoint in a protocol registered before the trial begins ties the authors' hands and forces them to present only one analysis as the "primary" finding. All other analyses are meant to be specified as either secondary or ad-hoc, thus addressing the concern that a selectively chosen subset of all conducted analyses are presented and given more than the appropriate weight in the discussion of results. No economics paper did this. We are aware of the fact that writing a protocol and registering it is now encouraged or required by groups such as JPAL[6], however this was not mentioned in any of the studies, no links or references to protocols were provided, and the rule linking adequacy of reporting to unclear risk of bias was applied. It is important to note that we enthusiastically support (and ourselves practice) the use of analyses conceived after a trial finishes in the formation of policy, but argue that they need to be described as such so that the reader knows how to weight the different types of evidence provided in the paper. The other issue at hand in reporting bias is that of even-handedness in presentation 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.

Imprecision

Only two economics papers attested to perform a prior sample size calculation. We are almost certain that some others did, (A. Banerjee et al. 2007; Parker 2010) but as none were reported, the economics literature failed to report adequately/be at low risk of bias almost categorically on this bias. We considered contacting authors to solicit such information, but decided against doing so in light of evidence that doing so would lead to biased responses (Haahr and Hróbjartsson 2006) and our rule tying inadequacy of reporting to risk of bias was applied.

Subgroup analyses

We analyze the bias assessments by a variety of subgroups, the results of which are shown in Appendix 3.3 and 3.4. We found that papers published more recently (e.g. in the 2010-2011 amendment to our initial search) did not do consistently better than their earlier-published counterparts. In medicine, we observe better reporting and lower risk of the six biases in the more recent set of RCT reports. This result is unsurprising given the increase over time in the awareness and acceptance of the CONSORT Statement guidelines and relevance of surrounding issues. Surprisingly, 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 strikingly similar to performance of papers in the other 47 economics journals we included. The only pattern

[6] See http://www.povertyactionlab.org/Hypothesis-Registry

we found was that papers reporting the results of economics RCTs taking place in developing countries did consistently worse than papers reporting the results of trials taking place in the US, Canada, and Europe. We find no such difference between those medical RCTs run in developed countries compared to those run in developing countries.

Section V: Ways Forward

We have presented evidence that RCTs in economics published between 2001 and 2011 did not utilize the large medical literature on bias minimization in the design and conduct of trials and, as a result, these trials are at unnecessary risk of bias in their analyses. Our work draws on a body of medical literature which has linked poor trial design, conduct, and reporting to exaggerated estimates of treatment effects. (David Moher et al. 1998; Schulz, Altman, and Moher 2010; Schulz, Altman, and Moher 2010) The identification of these shortcomings led to the systems of standards now used by medical trialists and journal editors which we draw upon for our grid. The establishment and acceptance of these standards in medicine has, in turn, led to an increase in the quality of articles reporting the results of trials. (Plint et al. 2006)

As discussed in Section III, the economics literature has begun to address several of these issues in the past few years. A recent exchange between two prominent economists touches on many such concerns and, despite their divergent views on other issues, the two authors agree on the fact that poor conduct of RCTs can bias interpretation. (Deaton 2009; Imbens 2010) A more thorough description of these concerns and other more practical problems of RCT implementation and interpretation is given in Duflo, Glennerster, and Kremer's article on how to conduct RCTs. (Duflo, Glennerster, and Kremer 2007) From the trials collected and analyzed here, however, there seems to be no public consensus on how to run an RCT in the social sciences. Furthermore, our analysis suggests that economists have not adopted many of the tools that medical trialists use for minimizing the risk of certain biases in their reports.

To ensure that the quality of evidence provided in economics articles reporting the results of RCTs is as high as possible, we propose that a system of reporting standards be established in economics similar to the CONSORT Statement guidelines widely accepted in the medical literature. These standards would give authors a tool to use on three fronts: one, in writing scholarly articles reporting the results of RCTs for publication in peer-reviewed journals, two, in the initial design of the studies themselves, and three, in performing meta-analyses and critical reviews such as this article. The crux of the argument in favor of such standards is twofold: one, that providing this information in trial reports enables readers to assess the quality of the evidence provided in each article, and two, that enforcing such standards encourages careful conduct of trials as well as thorough reporting, both ultimately leading to the creation of evidence with minimized risk of bias.

In terms of implementation, standards for trials in economics could well differ from those in medicine, perhaps in the admissibility of non-pre-specified endpoints, for example, given the sophisticated statistical and econometric tools often employed in robustness checks and sensitivity analysis. 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 trial quality. Duflo, Glennerster and Kremer's article outlines several issues that should be included in any set of guidelines, (Duflo, Glennerster, and Kremer 2007) but their treatment of the issues is not exhaustive. We strongly suggest that, at the very least, the following issues from the CONSORT Statement should be part of any set of guidelines for RCT design and reporting: a CONSORT-style diagram of flow of participants, a trial protocol registration system, which would include pre-specifying a primary analysis and providing explicit, sample size calculations conducted prior to trial entry and, in general, insistence on the intent-to-treat principle[7] for the primary analysis.

We also recognize that this is a field ripe for more analysis. Productive avenues of inquiry include mathematical simulation of the different types of biases to estimate how much the treatment effects in the literature to date should be discounted, investigation of publication bias in RCTs, and constructing a taxonomy of phases for trials in economics to help us know better when and how to apply the lessons from bias in medical trials. Additionally, though our initial investigation engaged with questions of external validity as well as internal, we have restricted our discussion here to internal validity to make our message more concise. External validity is arguably of similar importance and there is a rich literature on how to assess this in reports of RCTs. (Rothwell 2006) Each of these, however, is beyond the scope of this paper and we leave their pursuit to future research.

Lastly, we would like to mention that a major weakness of our study is the number of graders we used. Our grading 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. That said, we provide evidence for each assessment made and the differences we find between the two sets of RCT reports are so stark that it is unlikely to be solely the result of measurement error. We hope, as an extension of this project, to have a website which makes available the grid, our grades, and a database for others to enter their grades using the grid in order to refine the assessments presented here.

Section VI: Conclusion

In this study, we identified and discussed the potential for bias in the reports of randomized controlled trials in economics. From two of the main bias identification and minimization tools used by the medical literature, we crafted an evaluation tool, which we call the grid, to evaluate the adequacy of reporting and risk of six major biases in RCTs in economics. We evaluated a set of articles reporting the results of RCTs from 50 top economics journals and found that these articles performed poorly both in terms of providing the reader adequate information with which to assess the quality of the evidence provided by the study, and in terms of minimizing the risk of these six types of bias which have been associated with exaggerated treatment effects. We concluded by suggesting that the field of economics develop and adopt a set of reporting guidelines both to require the same degree of clarity and precision in the reports of RCTs that is demanded in medicine and to serve as a quality assessment tool to evaluate results that are published.

There are two main contributions of our analysis: methodological and empirical. In terms of methodology, we have discussed the nature of a set of biases and problems we believe RCTs are particularly prone to, catalogued the evidence of such problems skewing results in the medical literature, and provided a tool which can be used both to evaluate risk of bias in reports of RCTs as well as to assist in the design of future RCTs. Empirically, we showed that the reports of trials in economics published between 2000 and 20011 inadequately reported the risks of these bias according to the standards we derived from the medical literature, and that the design and implementation of many of these trials suggests they have made mistakes similar to those made in the past in the medical literature. Both findings suggest problems which have been associated with exaggerated treatment effects in

[7] Strict adherence to ITT without concurrent per-protocol analysis may not be advisable in non-inferiority trials. (Campbell, Elbourne, and Altman 2004; Piaggio et al. 2012)

the medical literature and raise serious concerns about the strength of the conclusions reached in some of the studies in economics scrutinized here.

Going forward, we hope that our study will lead to the establishment and acceptance of a set of standards for reporting RCTs that will minimize these biases in published reports of RCTs in the economics literature and will help readers to assess the quality of evidence provided in these reports. We hope it will also lead to increased efforts by trialists themselves to avoid these pitfalls in the design, execution, and analysis of their trials. Such efforts would lead to higher quality evidence and, we hope, the implementation of policy closer to the optimal.

References

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): 1064–1069.

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*: 1365–1372.

Banerjee, Abhijit. 2007. *Making Aid Work*. The MIT Press.

Bell, JA. 1948. “Pertussis Immunization; Use of Two Doses of an Alum-precipitated Mixture of Diphtheria Toxoid and Pertussis Vaccine.” *JAMA: The Journal of the American Medical Association* 137 (15): 1276.

Campbell, Marion K., Diana R. Elbourne, and Douglas G. Altman. 2004. “CONSORT Statement: Extension to Cluster Randomised Trials.” *British Medical Journal* 328 (7441): 702–708.

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): 448–452.

Collier, Roger. 2009. “Legumes, Lemons and Streptomycin: A Short History of the Clinical Trial.” *Canadian Medical Association Journal* 180 (1): 23–24.

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): 317–339.

Deaton, A. S. 2009. “Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development”. National Bureau of Economic Research.

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.

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

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): 818–822.

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

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): 359–370.

Imbens, Guido W. 2010. “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).” *Journal of Economic Literature*: 399–423.

Ioannidis, John PA, Anna-Bettina Haidich, Maroudia Pappa, Nikos Pantazis, Styliani I. Kokori, Maria G. Tektonidou, Despina G. Contopoulos-Ioannidis, and Joseph Lau. 2001. “Comparison of Evidence of Treatment Effects in Randomized and Nonrandomized Studies.” *JAMA: The Journal of the American Medical Association* 286 (7): 821–830.

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): 1054–1060.

Karlowski, Thomas R., Thomas C. Chalmers, Lawrence D. Frenkel, Albert Z. Kapikian, Thomas L. Lewis, and John M. Lynch. 1975. “Ascorbic Acid for the Common Cold.” *JAMA: The Journal of the American Medical Association* 231 (10): 1038–1042.

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.

Lewis, J. A., and D. Machin. 1993. “Intention to Treat–who Should Use ITT?” *British Journal of Cancer* 68 (4): 647.

Meldrum, Marcia L. 2000. “A Brief History of the Randomized Controlled Trial: From Oranges and Lemons to the Gold Standard.” *Hematology/oncology Clinics of North America* 14 (4): 745–760.

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: 657–662.

Moher, David, Ba’ Pham, Alison Jones, Deborah J. Cook, Alejandro R. Jadad, Michael Moher, Peter Tugwell, and Terry P. Klassen. 1998. “Does Quality of Reports of Randomised Trials Affect Estimates of Intervention Efficacy Reported in Meta-analyses?” *The Lancet* 352 (9128): 609–613.

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): 209–216.

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.

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.

Piaggio, G., D. R. Elbourne, S. J. Pocock, S. J. Evans, and D. G. Altman. 2012. "Reporting of Noninferiority and Equivalence Randomized Trials: Extension of the CONSORT 2010 Statement." *JAMA: The Journal of the American Medical Association* 308 (24): 2594–2604.

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): 408–412.

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.

Tannock, I. F. 1996. "False-positive Results in Clinical Trials: Multiple Significance Tests and the Problem of Unreported Comparisons." *JNCI Journal of the National Cancer Institute* 88 (3-4): 206.

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.

Twyman, Richard. 2004. "A Brief History of Clinical Trials." *The Human Genome* 22.

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

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*.

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.

Appendix 1: The Grid

Section: Selection Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
A.	□ **Randomisation generation and implementation** ○ Do the authors provide sufficient information that the reader can assess the methods used to generate the random allocation sequence and the likelihood of bias in treatment allocation? ○ Does the paper explain who generated the allocation sequence, who enrolled participants and who assigned participants to the trial group?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Flow of participants -** does the paper state how many participants: ○ Were assessed for eligibility ○ Were eligible ○ Were enrolled ○ Were excluded ○ Were randomised to each intervention? ○ Are these numbers given in a clear, easily interpretable manner?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Baseline demographics -** are the study groups compared at the baseline for important demographic and clinical characteristics, allowing the reader to assess how comparable they are?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Performance Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
B.	□ **Blinding and data collection –** participants are ideally blinded to their allocation status. Are the participants in the trial blinded? If participants are not blinded, are the study endpoints objective and collected by someone unlikely to influence the response differentially? (e.g. not data from self-reporting or someone affiliated with the intervention) If not, does the paper discuss the resultant risk of bias and what is done to control for it?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Blinding and participant conduct** – again, participants are ideally blinded to their allocation status. Does the paper mention whether blinding recipients was possible and, if so, considered? If not, does it discuss the potential problems from participants seeking care differentially as a result of being aware of their treatment allocation and whether these problems are likely to have occurred?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Detection Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
C.	□ **Data collection -** does the paper state: ○ How the data is collected ○ Who is collecting the data ○ What relationship, if any, the data collectors have to the intervention? ○ Does the paper mention whether blinding data collectors was possible and, if so, considered?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Attrition Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
D.	□ **Flow of participants -** does the paper state how many participants: ○ Received each intervention ○ Did not receive each intervention ○ Were followed up ○ Were lost to follow up ○ Were included for analysis ○ Were excluded from the analysis by the investigators?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Number of participants/intention to treat -** 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?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Reporting Bias		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
E.	□ **Pre-specified protocol and analysis plan -** does the paper have a pre-specified protocol and analysis plan for conduct and evaluation of the trial?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Outcomes and summary of results** ○ Are all presented outcomes defined as primary, secondary or exploratory? ○ Are the results presented for all planned primary and secondary endpoints? ○ Are the results presented in an intuitive manner, including the summary of each outcome and the measured effect size with a confidence interval?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Reporting Bias (cont’d)		Reported adequately?		Low risk of bias?	
	Issue	Judgment	Description	Judgment	Description
E.	□ **Ancillary analyses –** do the authors present or offer a link to an appendix listing the exploratory analyses performed but not presented in the paper?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:
	□ **Interpretation -** does the interpretation of the results: ○ Offer a synopsis of the findings ○ Provide a consideration of possible mechanisms and explanations ○ Offer comparison with relevant findings from other studies and discuss the results of the trial in the context of existing evidence, evidence which is not limited to evidence that supports the results of the current trial ○ Discuss limitations of the present study ○ Exercise special care when evaluating the results of a trial with multiple comparisons (e.g. multiple endpoints or subgroup analyses)?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Section: Sample Size		Reported adequately?		Low risk of imprecision?	
	Issue	Judgment	Description	Judgment	Description
F.	□ **Sample size -** do the authors indicate whether they conduct a sample size calculation and if so, how?	Yes No	Quote: Comment:	Yes No / Unclear	Quote: Comment:

Appendix 2: Articles Evaluated in the Analysis

Articles in economics			
First Author	Journal	Year	Title
Anderson	Quarterly Journal of Economics	2010	Price Stickiness and Customer Antagonism
Angrist	American Economic Journal: Applied Economics	2009	Incentives and Services for College Achievement - Evidence from a Randomized Trial
Angrist	American Economic Review	2009	The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial
Ashenfelter	Journal of Econometrics	2005	Do Unemployment Insurance Recipients Actively Seek Work? Evidence from Randomized Trials in Four U.S. States
Ashraf	Quarterly Journal of Economics	2006	Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines
Attanasio	American Economic Journal: Applied Economics	2011	Subsidizing Vocational Training for Disadvantaged Youth in Colombia: Evidence from a Randomized Trial
Banerjee	American Economic Journal: Applied Economics	2010	Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India
Banerjee	Quarterly Journal of Economics	2007	Remedying Education: Evidence from Two Randomized Experiments in India
Barrera-Osorio	American Economic Journal: Applied Economics	2011	Improving the Design of Conditional Transfer Programs: Evidence from a Randomized Education Experiment in Colombia
Barrow	American Economic Journal: Economic Policy	2009	Technology's Edge: The Educational Benefits of Computer-Aided Instruction
Bertrand	Quarterly Journal of Economics	2010	What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment
Bjorkman	Quarterly Journal of Economics	2009	Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda
Blau	American Economic Review	2010	Can Mentoring Help Female Assistant Professors? Interim Results from a Randomized Trial
Bobonis	Journal of Human Resources	2006	Anemia and School Participation
Cai	American Economic Review	2009	Observational Learning: Evidence from a Randomized Natural Field Experiment

Cohen	Quarterly Journal of Economics	2010	Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment
de Janvry	Journal of Development Economics	2010	The Supply- and Demand-Side Impacts of Credit Market Information
de Janvry	Journal of Economic Behavior and Organization	2010	Short on Shots: Are Calls for Cooperative Restraint Effective in Managing a Flu Vaccines Shortage?
de Mel	Quarterly Journal of Economics	2008	Returns to Capital in Microenterprises: Evidence from a Field Experiment
Duflo	American Economic Review	2011	Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya
Duflo	Quarterly Journal of Economics	2006	Saving Incentives for Low- and Middle-Income Families: Evidence from a Field Experiment with H&R Block
Duflo	Quarterly Journal of Economics	2003	The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment
Dupas	American Economic Journal: Applied Economics	2011	Do Teenagers Respond to HIV Risk Information? Evidence from a Field Experiment in Kenya
Fehr	American Economic Review	2007	Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment
Ferraro	American Economic Review	2011	The Persistence of Treatment Effects with Norm-Based Policy Instruments: Evidence from a Randomized Environmental Policy Experiment
Fryer	Quarterly Journal of Economics	2011	Financial Incentives and Student Achievement: Evidence from Randomized Trials
Gine	Journal of Development Economics	2009	Insurance, Credit, and Technology Adoption: Field Experimental Evidence from Malawi Gine, Xavier; Yang,
Glewwe	American Economic Journal: Applied Economics	2010	Teacher Incentives
Glewwe	American Economic Journal: Applied Economics	2009	Many Children Left Behind? Textbooks and Test Scores in Kenya
Glewwe	Journal of Development Economics	2004	Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya
Harrison	Journal of Economic Behavior and Organization	2009	Risk Attitudes, Randomization to Treatment, and Self-Selection into Experiments

Hu	Journal of Human Resources	2003	Marriage and Economic Incentives: Evidence from a Welfare Experiment
Huysentruyt	American Economic Journal: Applied Economics	2010	Child Benefit Support and Method of Payment: Evidence from a Randomized Experiment in Belgium
Karlan	Review of Economics and Statistics	2011	Teaching Entrepreneurship: Impact of Business Training on Microfinance Clients and Institutions
Karlan	Review of Financial Studies	2010	Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts
Karlan	American Economic Review	2008	Credit Elasticities in Less-Developed Economies: Implications for Microfinance,
Katz	Quarterly Journal of Economics	2001	Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment
Kleven	Econometrica	2011	Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark
Kremer	Quarterly Journal of Economics	2011	Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions
Kremer	Quarterly Journal of Economics	2007	The Illusion of Sustainability
Kremer	Review of Economics and Statistics	2009	Incentives to Learn
Linnemayr	Journal of Development Economics	2011	Almost Random: Evaluating a Large-Scale Randomized Nutrition Program in the Presence of Crossover
Michalopoulos	Journal of Public Economics	2005	When Financial Work Incentives Pay for Themselves: Evidence from a Randomized Social Experiment for Welfare Recipients
Miguel	Econometrica	2004	Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities
Muralidharan	Journal of Political Economy	2011	Teacher Performance Pay: Experimental Evidence from India
Olken	Journal of Political Economy	2007	Monitoring Corruption: evidence from a Field Experiment in Indonesia
Oster	American Economic Journal: Applied Economics	2011	Menstruation, Sanitary Products, and School Attendance: Evidence from a Randomized Evaluation
Pozo	American Economic Review	2006	Requiring a Math Skills Unit: Results of a Randomized Experiment
Rosholm	Journal of Applied Econometrics	2009	Is Labour Market Training a Curse for the Unemployed? Evidence from a Social Experiment

Saez	American Economic Journal: Economic Policy	2009	Details Matter: The Impact of Presentation and Information on the Take-up of Financial Incentives for Retirement Saving
Schady	Economics Letters	2008	Are Cash Transfers Made to Women Spent Like Other Sources of Income?
Schultz	Journal of Development Economics	2004	School Subsidies for the Poor: Evaluating the Mexican Progresa Poverty Program
Thornton	American Economic Review	2008	The Demand for, and Impact of, Learning HIV Status
van den Berg	International Economic Review	2006	Counseling and Monitoring of Unemployed Workers: Theory and Evidence from a Controlled Social Experiment

Articles in medicine

First Author	Journal	Year	Title
Albert	Journal of the American Medical Association	2001	Effect of Statin Therapy on C-Reactive Protein Levels: The Pravastatin Inflammation/CRP Evaluation (PRINCE): A Randomized Trial and Cohort Study
American Lung Association Asthma Clinical Research Centers	New England Journal of Medicine	2009	Efficacy of Esomeprazole for Treatment of Poorly Controlled Asthma
Aufderheide	New England Journal of Medicine	2011	A Trial of an Impedance Threshold Device in Out-of-Hospital Cardiac Arrest
Barwell	The Lancet	2004	Comparison of surgery and compression with compression alone in chronic venous ulceration (ESCHAR study): randomised controlled trial
Blanc	New England Journal of Medicine	2011	Earlier versus Later Start of Antiretroviral Therapy in HIV-Infected Adults with Tuberculosis
Blankensteijn	New England Journal of Medicine	2005	Two-Year outcomes after Conventional or Endovascular Repair of Abdominal Aortic Aneurysms
Church	Journal of the American Medical Association	2010	Effects of Aerobic and Resistance Training on Hemoglobin A1c Levels in Patients With Type 2 Diabetes
Cicardi	New England Journal of Medicine	2010	Ecallantide for the Treatment of Acute Attacks in Hereditary Angioedema
Conroy	New England Journal of Medicine	2011	FOLFIRINOX versus Gemcitabine for Metastatic Pancreatic Cancer
Cummings	New England Journal of Medicine	2010	Lasofoxifene in Postmenopausal Women with Osteoporosis
Cutland	The Lancet	2009	Chlorhexidine maternal-vaginal and neonate body wipes in sepsis and vertical transmission of pathogenic bacteria in South Africa: a randomised, controlled trial
de Smet	New England Journal of Medicine	2009	Decontamination of the Digestive Track and Oropharynx in ICU Patients
Decousus	New England Journal of Medicine	2010	Fondaparinux for the treatment of superficial-vein thrombosis in the legs
Dobscha	Journal of the American Medical Association	2009	Collaborative Care for Chronic Pain in Primary Care: A Cluster Randomized Trial

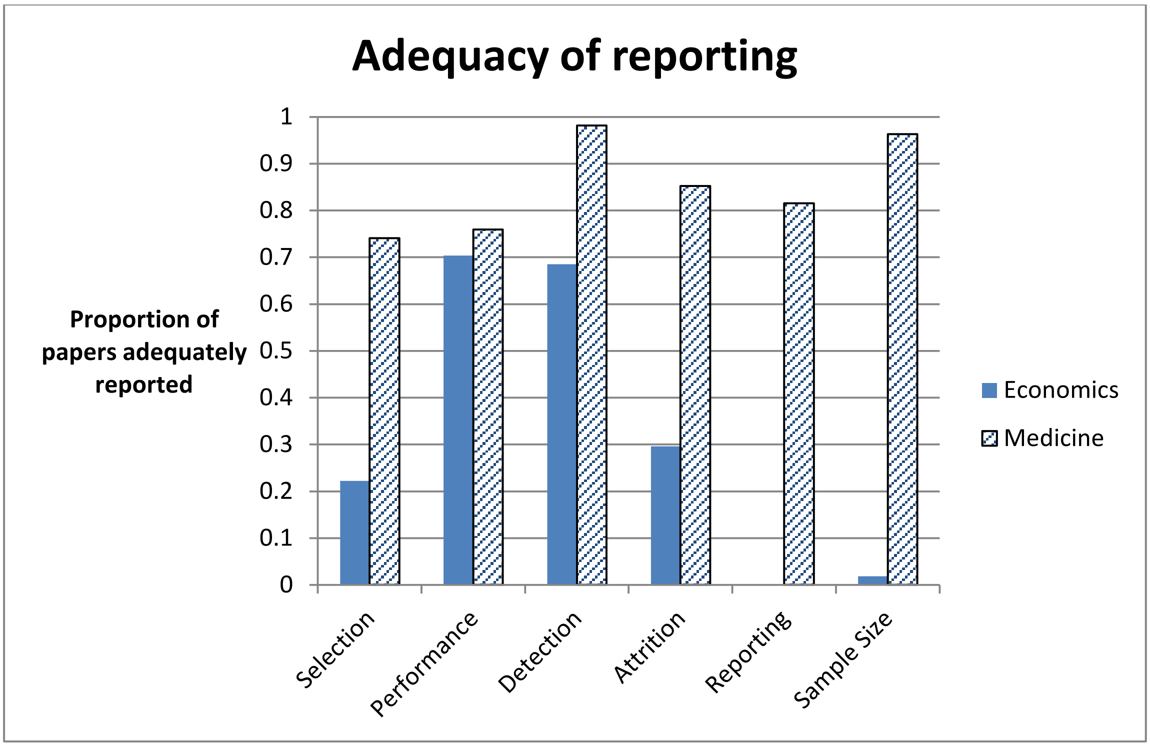
Dorsey	Journal of the American Medical Association	2007	Combination Therapy for Uncomplicated Falciparum Malaria in Ugandan Children: A Randomized Trial
Fergusson	New England Journal of Medicine	2008	A Comparison of Aprotinin and Lysine Analogues in High-Risk Cardiac Surgery
Glauser	New England Journal of Medicine	2010	Ethosuximide, Valproic Acid, and Lamotrigine in Childhood Absence Epilepsy
Gorelick	Journal of the American Medical Association	2003	Aspirin and Ticlopidine for Prevention of Recurrent Stroke in Black Patients: A Randomized Trial
Herbst	The Lancet	2011	Efficacy of bevacizumab plus erlotinib versus erlotinib alone in advanced non-small-cell lung cancer after failure of standard first-line chemotherapy (BeTa): a double-blind, placebo-controlled, phase 3 trial
Karunajeewa	New England Journal of Medicine	2008	A Trial of Combination Antimalarial Therapy in Children from Papua New Guinea
Kawamori	The Lancet	2009	Voglibose for prevention of type 2 diabetes mellitus: a randomised, double-blind trial in Japanese individuals with impaired glucose tolerance
Koopmans	The Lancet	2009	Induction of labour versus expectant monitoring for gestational hypertension or mild pre-eclampsia after 36 weeks' gestation (HYPITAT): a multicentre, open-label randomised controlled trial
Krueger	New England Journal of Medicine	2007	A Human Interleukin-12/23 Monoclonal Antibody for the Treatment of Psoriasis
Lamb	The Lancet	2010	Group cognitive behavioural treatment for low-back pain in primary care: a randomised controlled trial and cost-effectiveness analysis
Lazcano-Ponce	The Lancet	2011	Self-collection of vaginal specimens for human papillomavirus testing in cervical cancer prevention (MARCH): a community-based randomised controlled trial
Lemanske	New England Journal of Medicine	2010	Step-up Therapy for Children with Uncontrolled Asthma Receiving Inhaled Corticosteroids
Lennox	The Lancet	2009	Safety and efficacy of raltegravir-based versus efavirenz-based combination therapy in treatment-naïve patients with HIV-1 infection: a multicentre, double-blind randomised controlled trial
Lenze	Journal of the American Medical Association	2009	Escitalopram for Older Adults With Generalized Anxiety Disorder

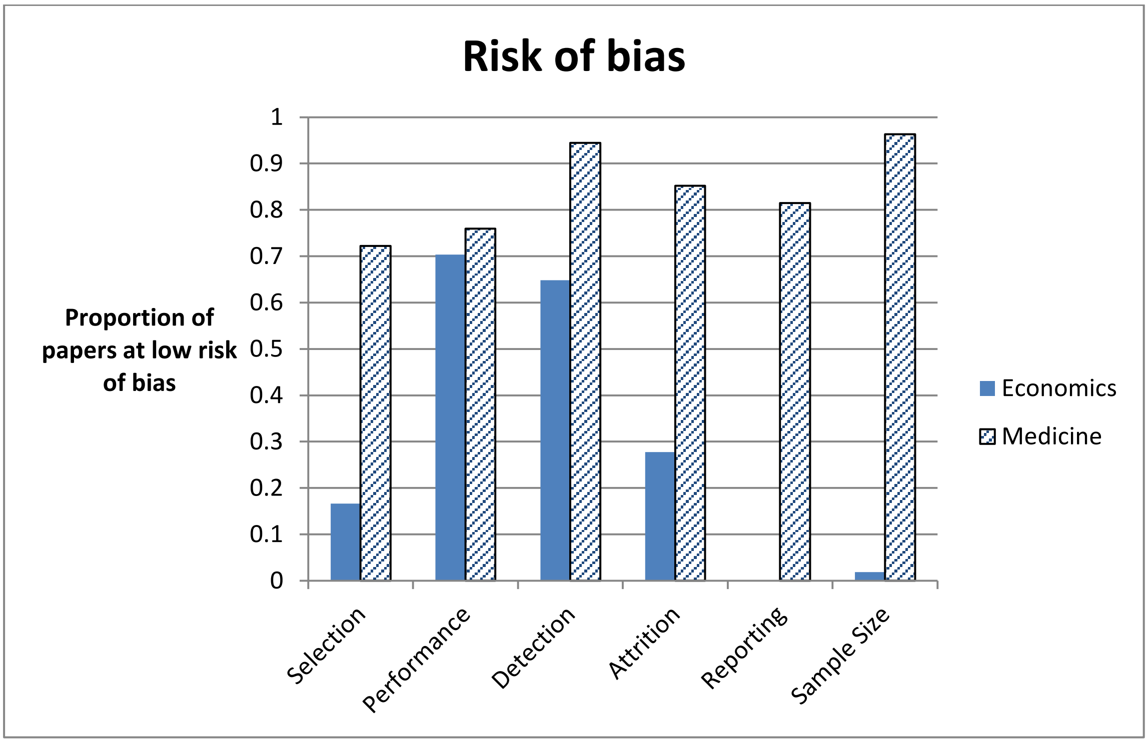
McFall	Journal of the American Medical Association	2010	Integrating Tobacco Cessation Into Mental Health Care for Posttraumatic Stress Disorder A Randomized Controlled Trial
Montalescot	Journal of the American Medical Association	2009	Immediate vs Delayed Intervention for Acute Coronary Syndromes: A Randomized Clinical Trial
National Lung Screening Trial Research Team	New England Journal of Medicine	2011	Reduced lung-cancer mortality with low-dose computed tomographic screening.
Navarra	The Lancet	2011	Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial
Nissen	New England Journal of Medicine	2006	Effect of ACAT Inhibition on the Progression of Coronary Atherosclerosis
Papanikolaou	New England Journal of Medicine	2006	In Vitro Fertilization with Single Blastocyst-Stage versus Single Cleavage-Stage Embryos
Peikes	Journal of the American Medical Association	2009	Effects of Care Coordination on Hospitalization, Quality of Care, and Health Care Expenditures Among Medicare Beneficiaries
Perondi	New England Journal of Medicine	2004	A Comparison of High-Dose and Standard-Dose Epinephrine in Children with Cardiac Arrest
Pichichero	Journal of the American Medical Association	2005	Combined Tetanus, Diphtheria, and 5-Component Pertussis Vaccine for use in Adolescents and Adults
Pimentel	New England Journal of Medicine	2011	Rifaximin therapy for patients with irritable bowel syndrome without constipation.
Riddler	New England Journal of Medicine	2008	Class-Sparing Regimens for Initial Treatment of HIV-1 Infection
Sandler	New England Journal of Medicine	2006	Paclitaxel-Carboplatin Alone or with Bevacizumab for Non-Small-Cell Lung Cancer
Sandset	The Lancet	2011	The angiotensin-receptor blocker candesartan for treatment of acute stroke (SCAST): a randomised, placebo-controlled, double-blind trial.
Scolnik	Journal of the American Medical Association	2006	Controlled Delivery of High vs Low Humidity vs Mist Therapy for Croup in Emergency Departments: A Randomized Controlled Trial
Staessen	Journal of the American Medical Association	2004	Antihypertensive Treatment Based on Blood Pressure Measurement at Home or in the Physician's Office: A Randomized Controlled Trial

Tardif	The Lancet	2008	Effects of succinobucol (AGI-1067) after an acute coronary syndrome: a randomised, double-blind, placebo controlled trial
Tate	Journal of the American Medical Association	2003	Effects of Internet Behavioral Counseling on Weight Loss in Adults at Risk for Type 2 Diabetes: A Randomized Trial
Tonetti	New England Journal of Medicine	2007	Treatment of Periodontitis and Endothelial Function
Tylleskär	The Lancet	2011	Exclusive breastfeeding promotion by peer counsellors in sub-Saharan Africa (PROMISE-EBF): a cluster-randomised trial.
Van den Berghe	New England Journal of Medicine	2006	Intensive Insulin Therapy in the Medical ICU
van Ruler	Journal of the American Medical Association	2007	Comparison of On-Demand vs Planned Relaparotomy Strategy in Patients With Severe Peritonitis: A Randomized Trial
Vollenhoben	The Lancet	2009	Addition of infliximab compared with addition of sulfasalazine and hydroxychloroquine to methotrexate in patients with early rheumatoid arthritis (Swefot trial): 1-year results of a randomised trial
Wainwright	Journal of the American Medical Association	2011	Effect of bronchoalveolar lavage-directed therapy on Pseudomonas aeruginosa infection and structural lung injury in children with cystic fibrosis: a randomized trial.
Walton	Journal of the American Medical Association	2010	Effects of a Brief Intervention for Reducing Violence and Alcohol Misuse Among Adolescents
Wilkens	Journal of the American Medical Association	2010	Effect of Glucosamine on Pain-Related Disability in Patients With Chronic Low Back Pain and Degenerative Lumbar Osteoarthritis
Zeuzem	New England Journal of Medicine	2011	Telaprevir for retreatment of HCV infection

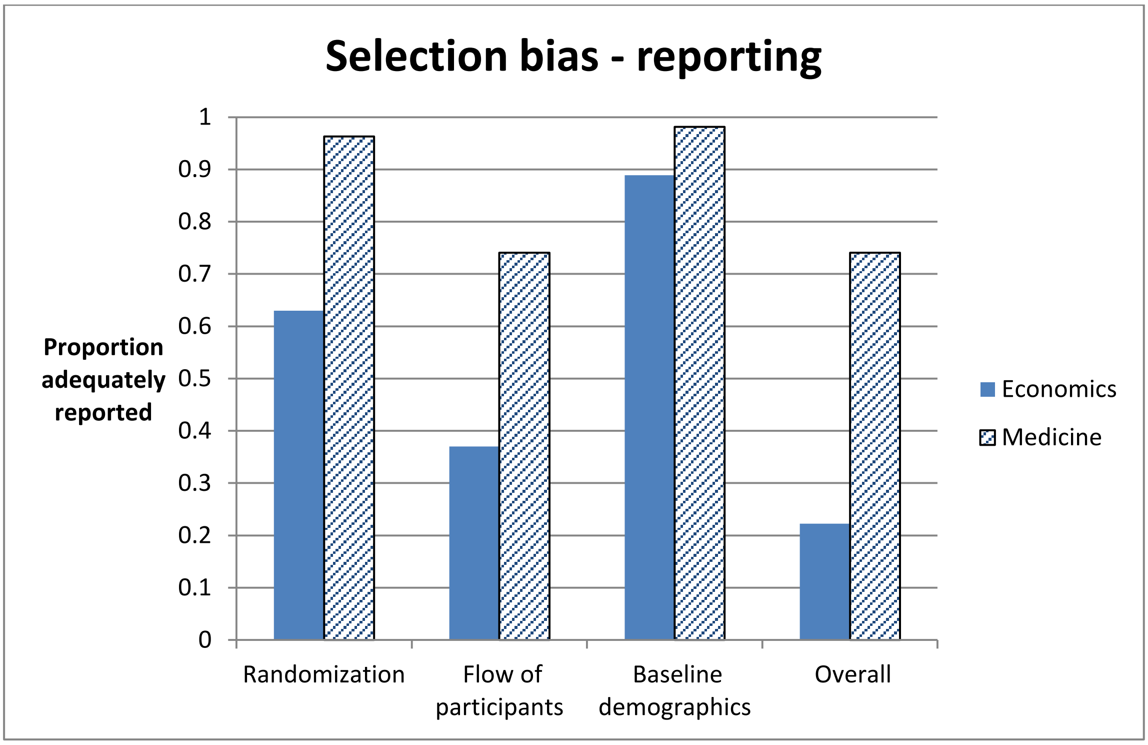
Appendix 3: Bar Charts with Grades

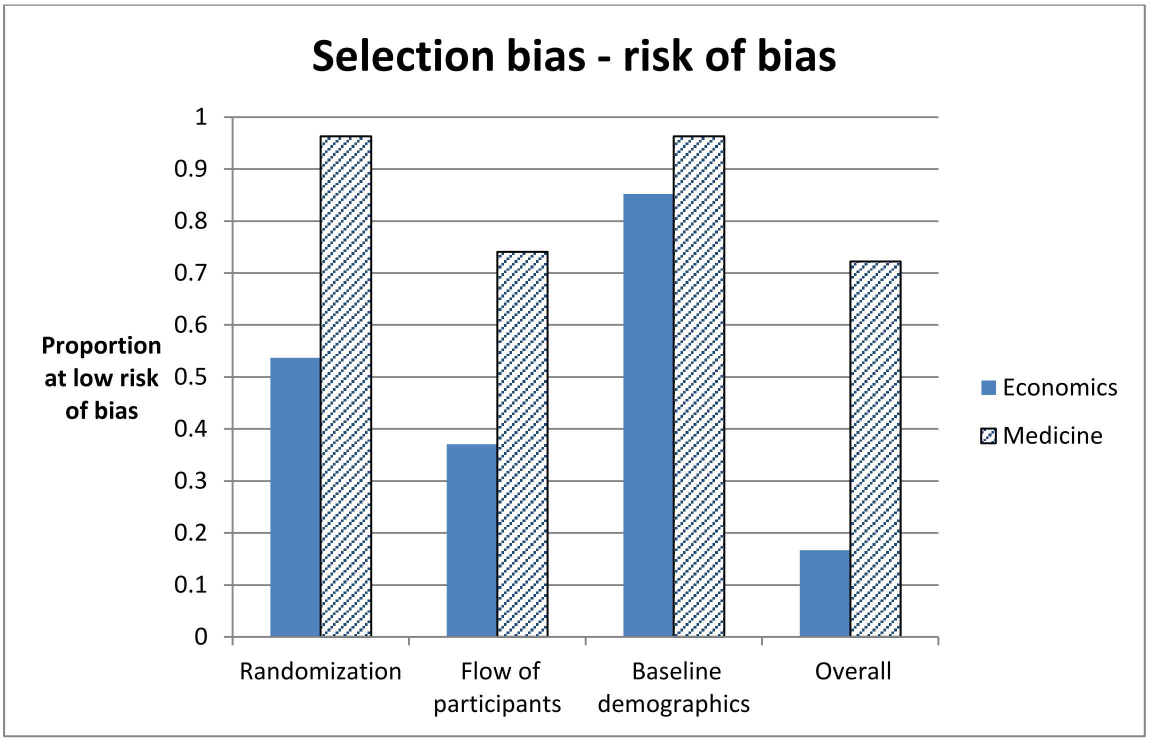
Appendix 3.1: Overall performance

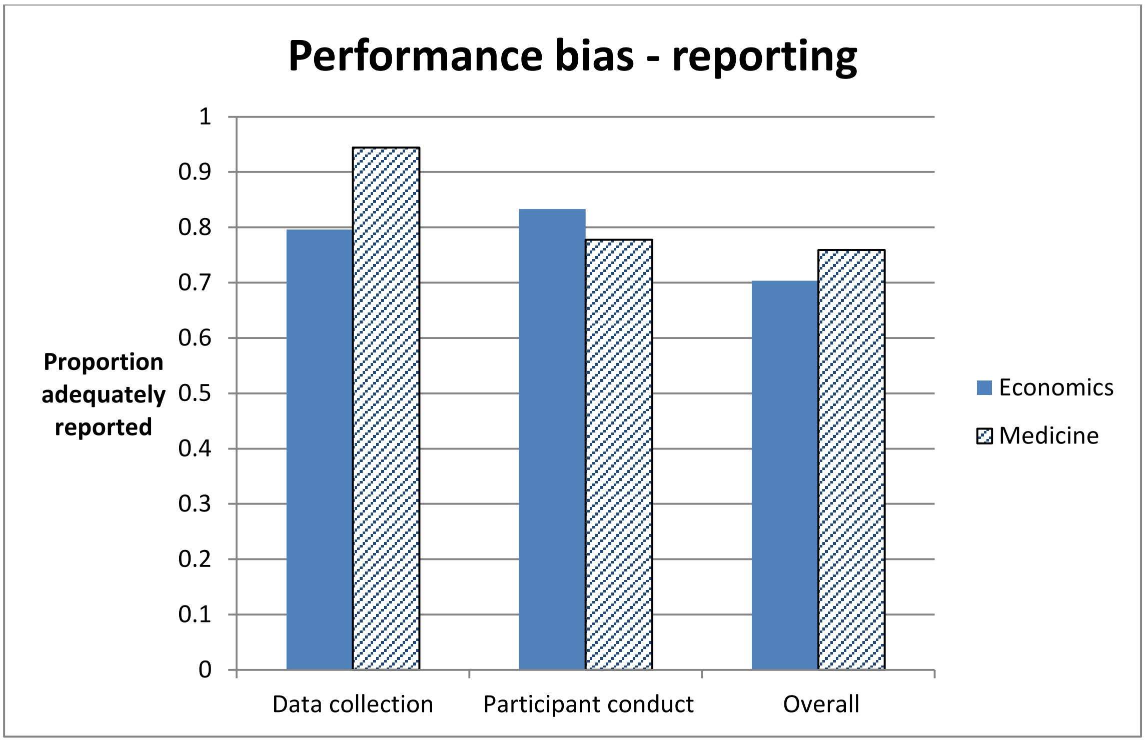


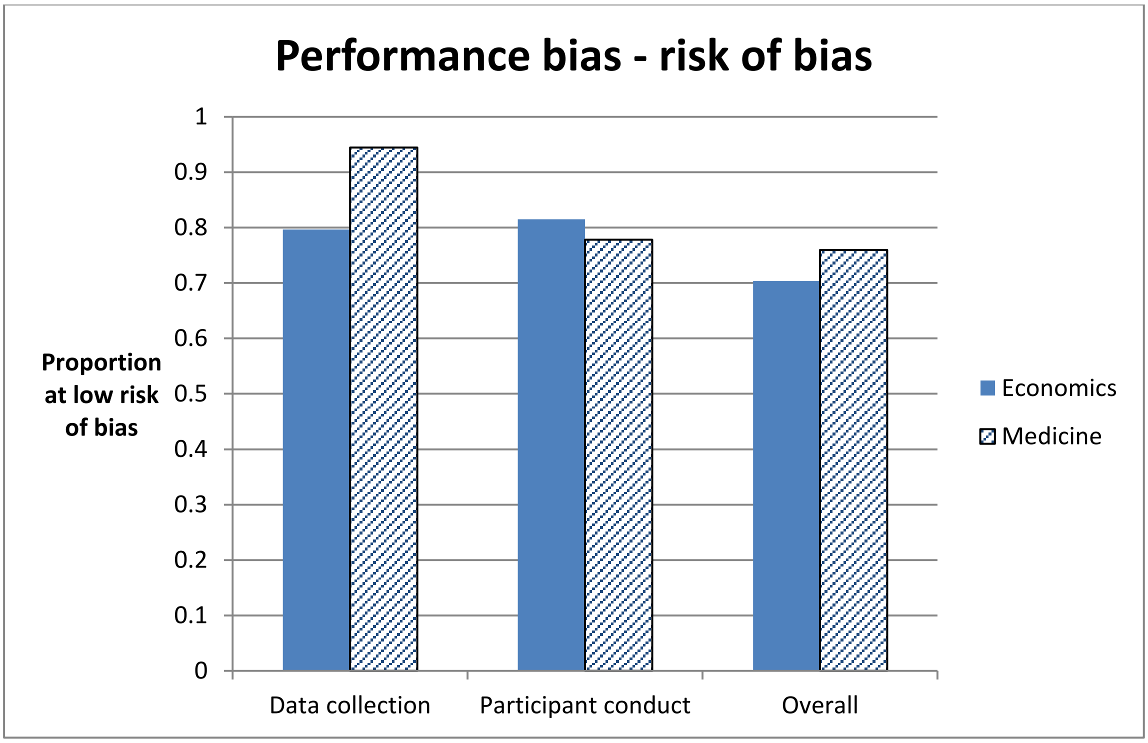


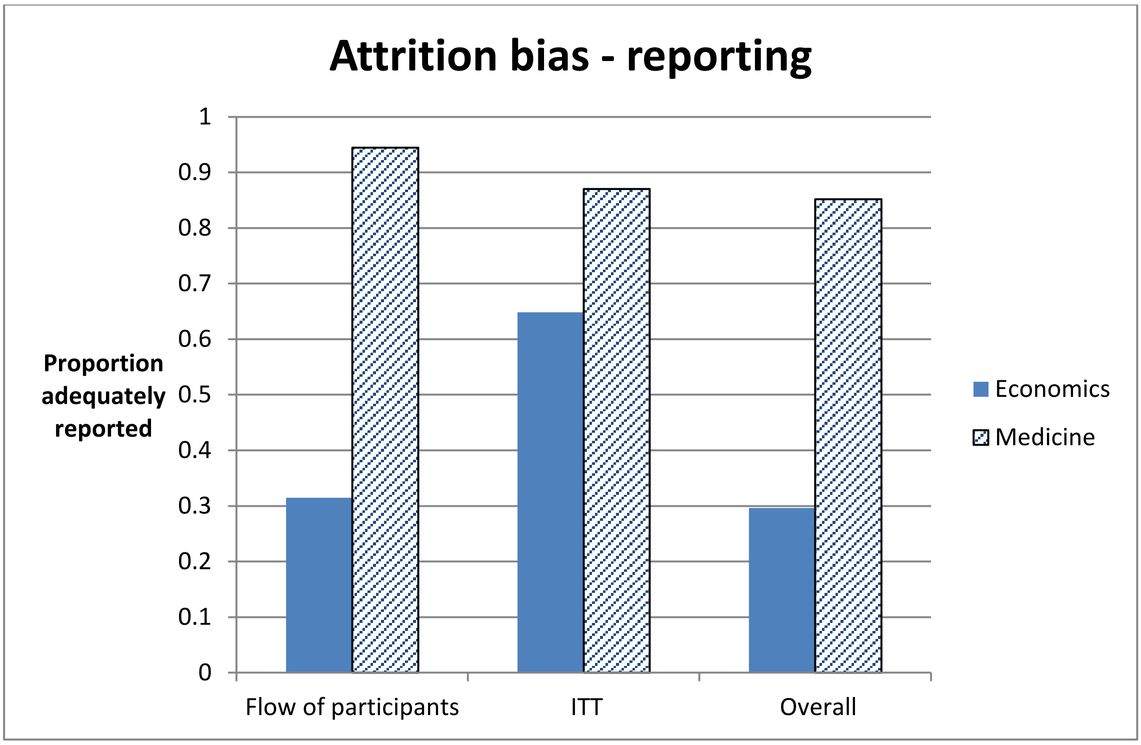
Appendix 3.2: Performance broken down by sub-issue

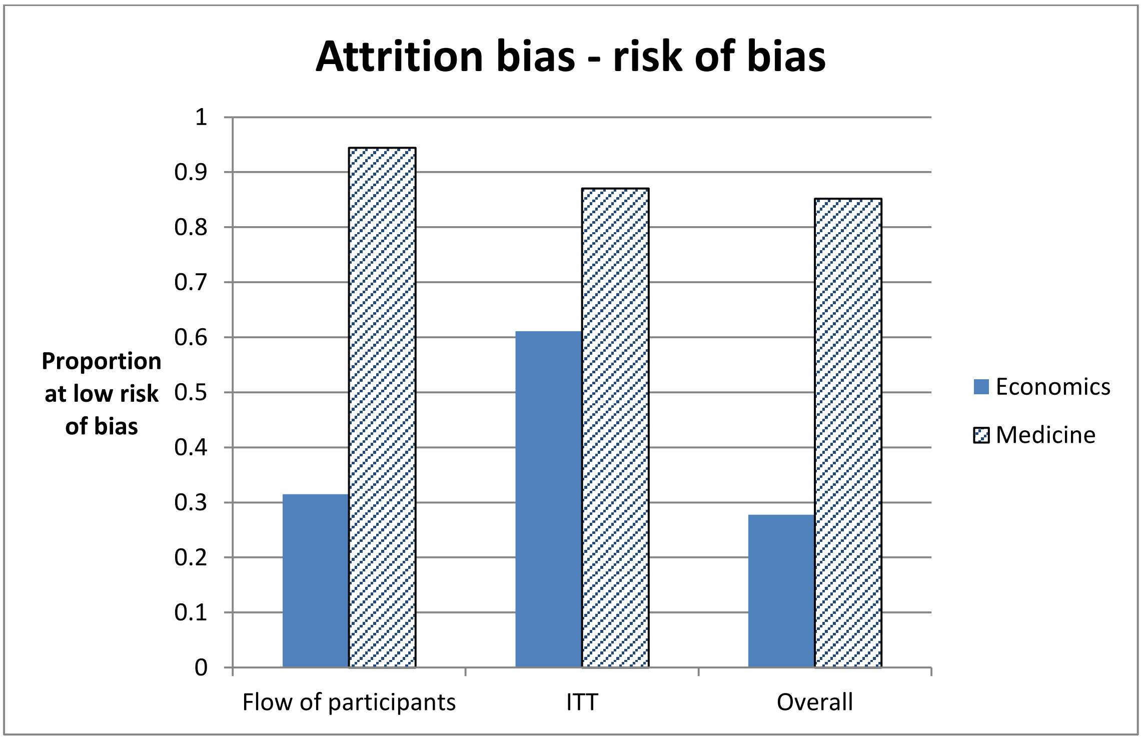


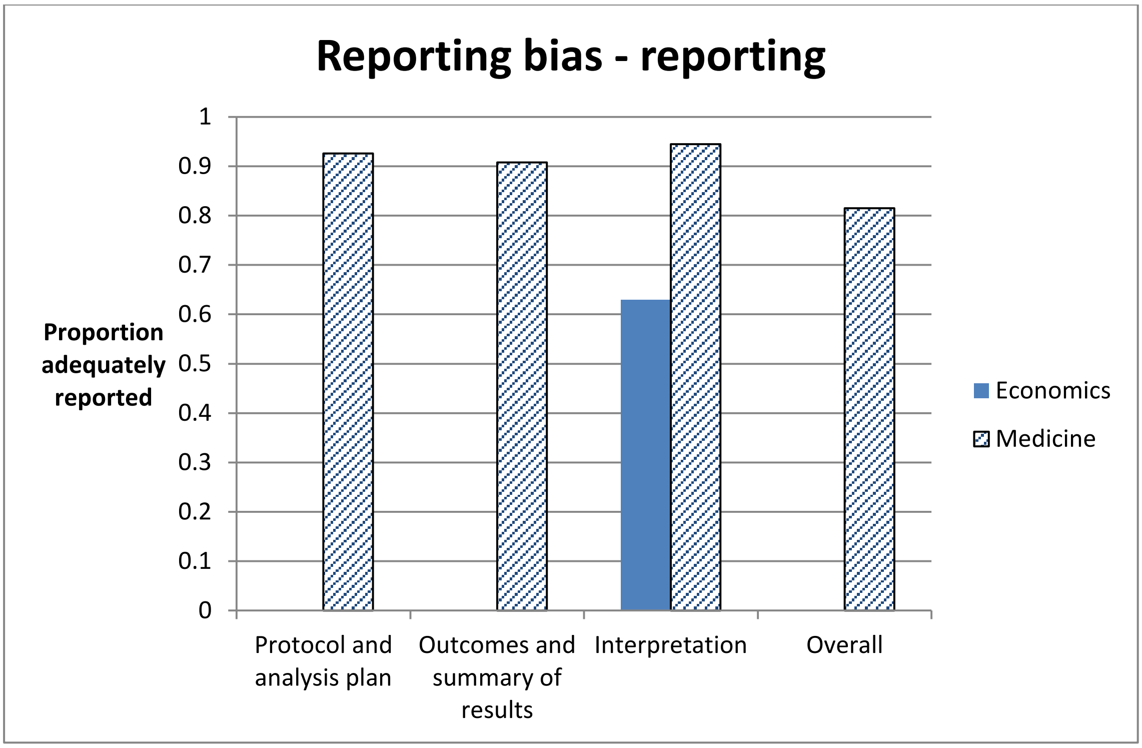


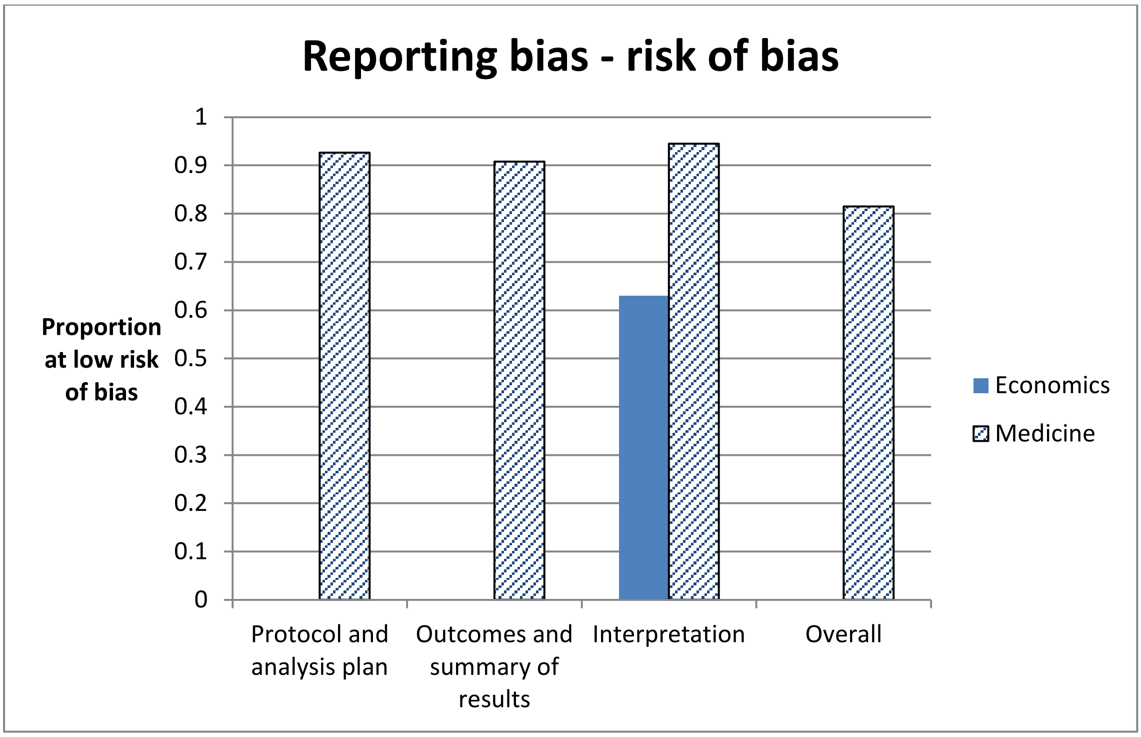

Performance bias - reporting
Proportion adequately reported
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Data collection
Participant conduct
Overall
Economics
Medicine

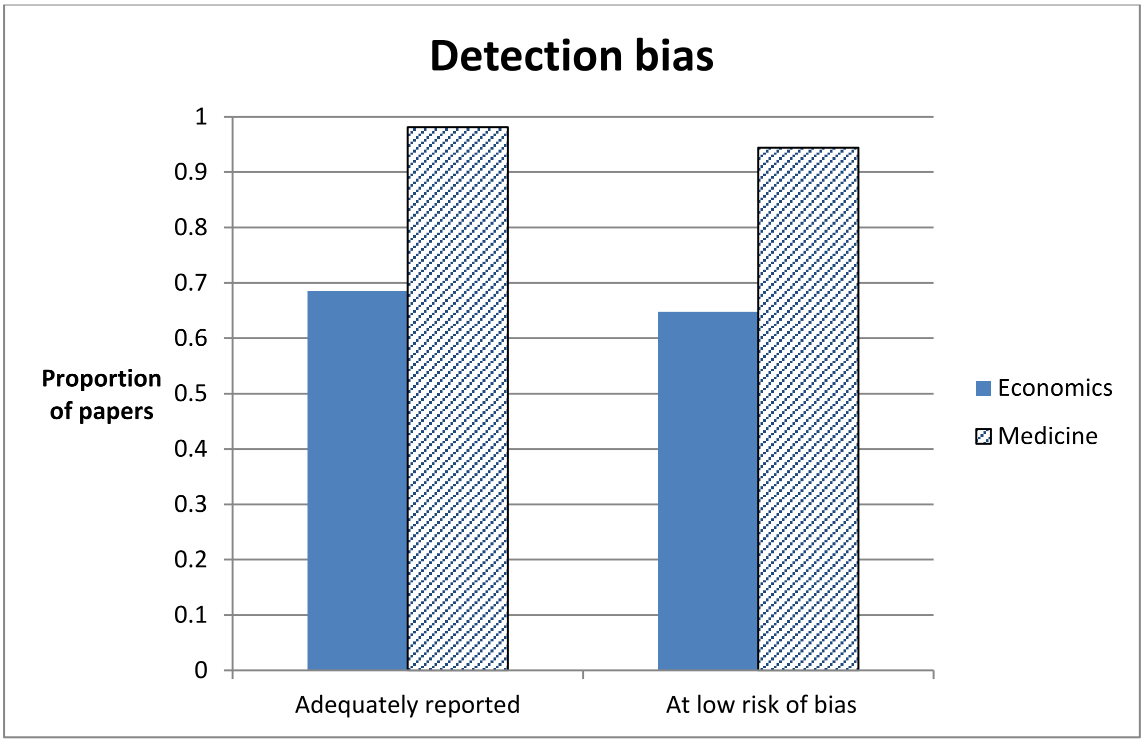

Performance bias - risk of bias
Proportion at low risk of bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Data collection
Participant conduct
Overall
Economics
Medicine

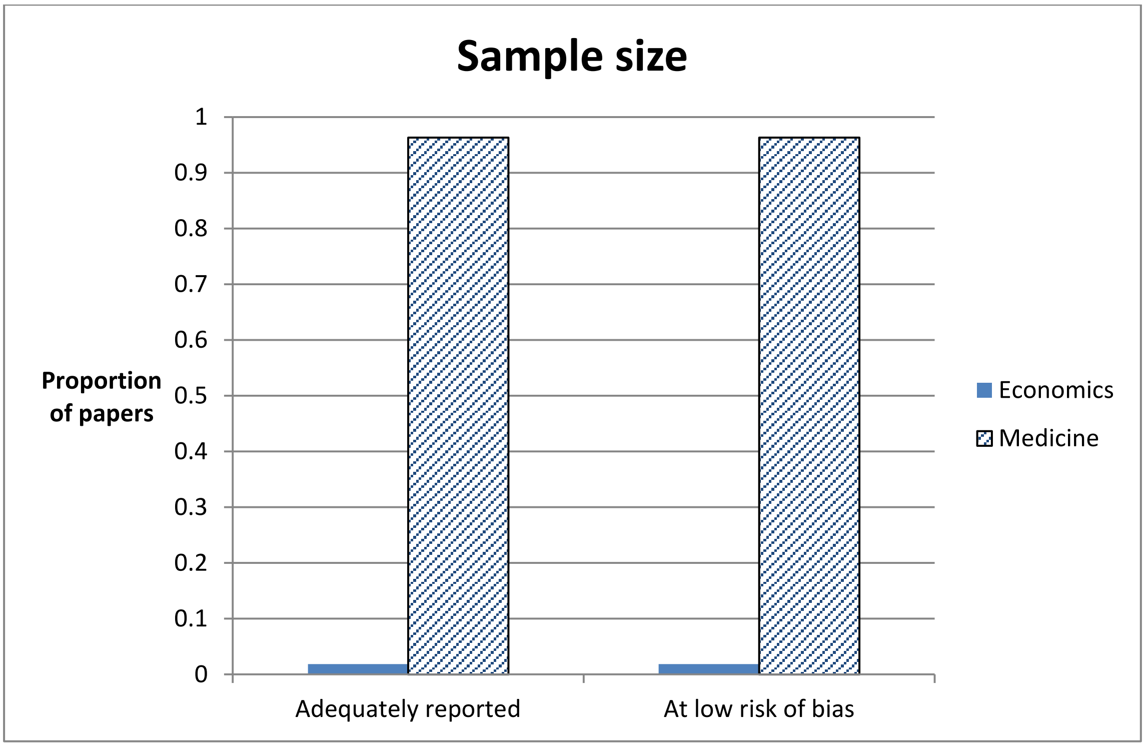

Attrition bias - reporting
Proportion adequately reported
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Flow of participants
ITT
Overall
Economics
Medicine


Attrition bias - risk of bias
Proportion at low risk of bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Flow of participants
ITT
Overall
Economics
Medicine

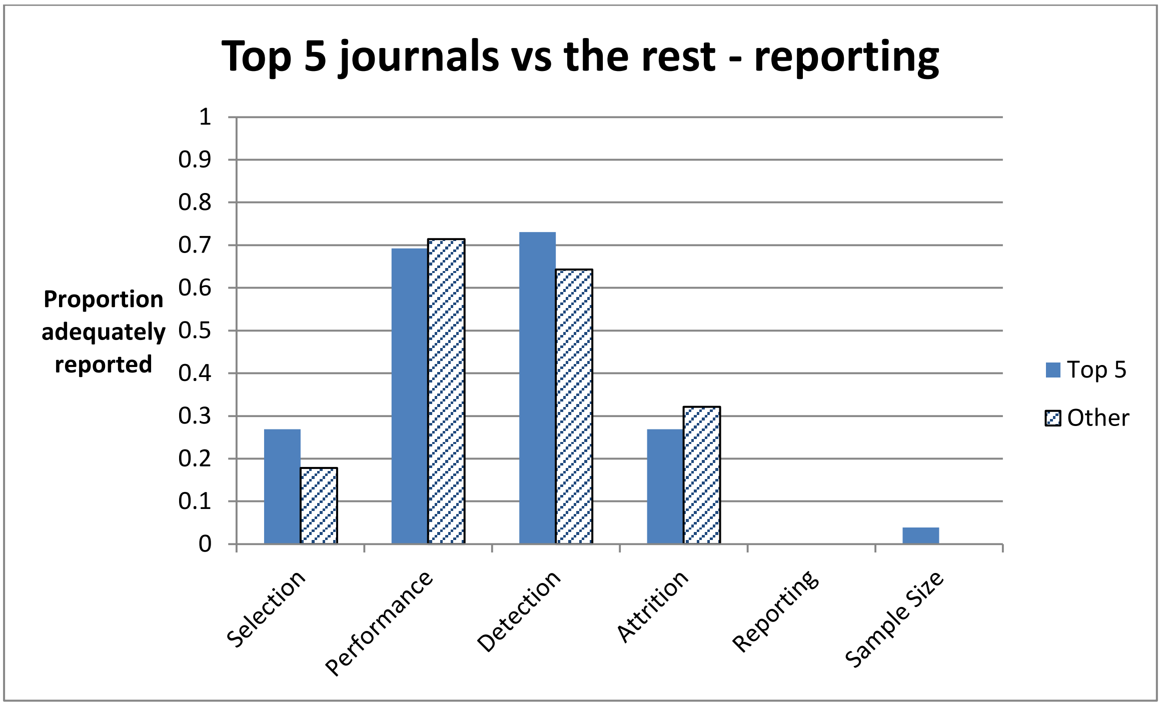

Reporting bias - reporting
Proportion adequately reported
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Protocol and analysis plan
Outcomes and summary of results
Interpretation
Overall
Economics
Medicine

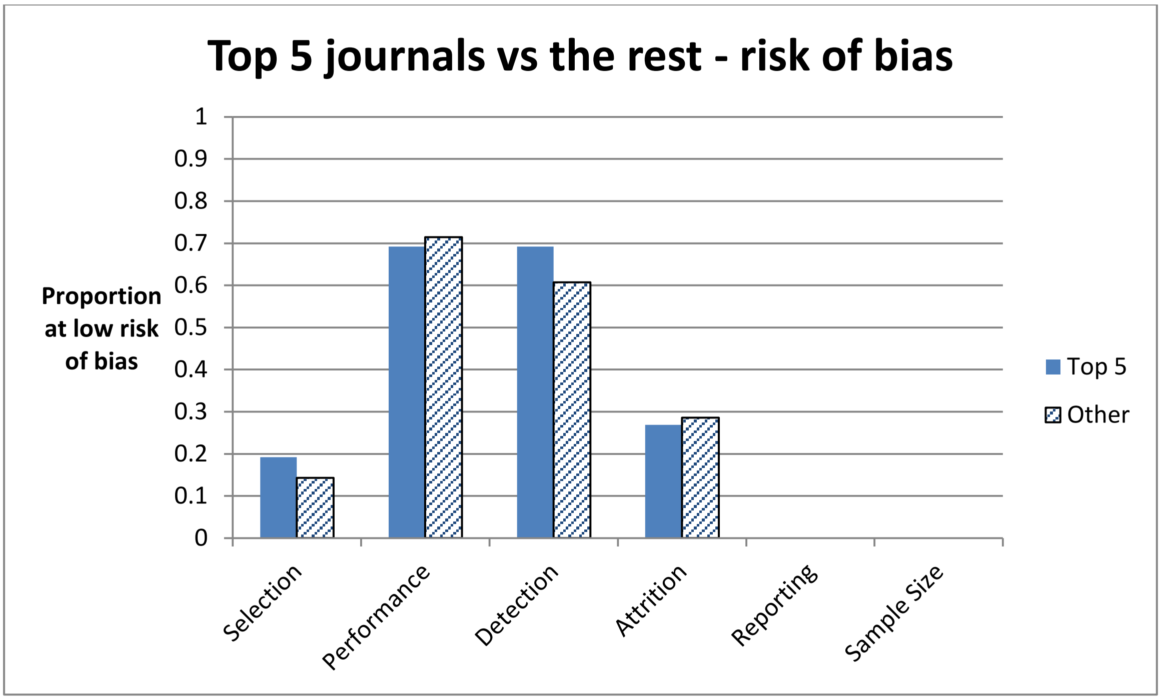

Reporting bias - risk of bias
Proportion at low risk of bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Protocol and analysis plan
Outcomes and summary of results
Interpretation
Overall
Economics
Medicine


Detection bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion of papers
Adequately reported
At low risk of bias
Economics
Medicine

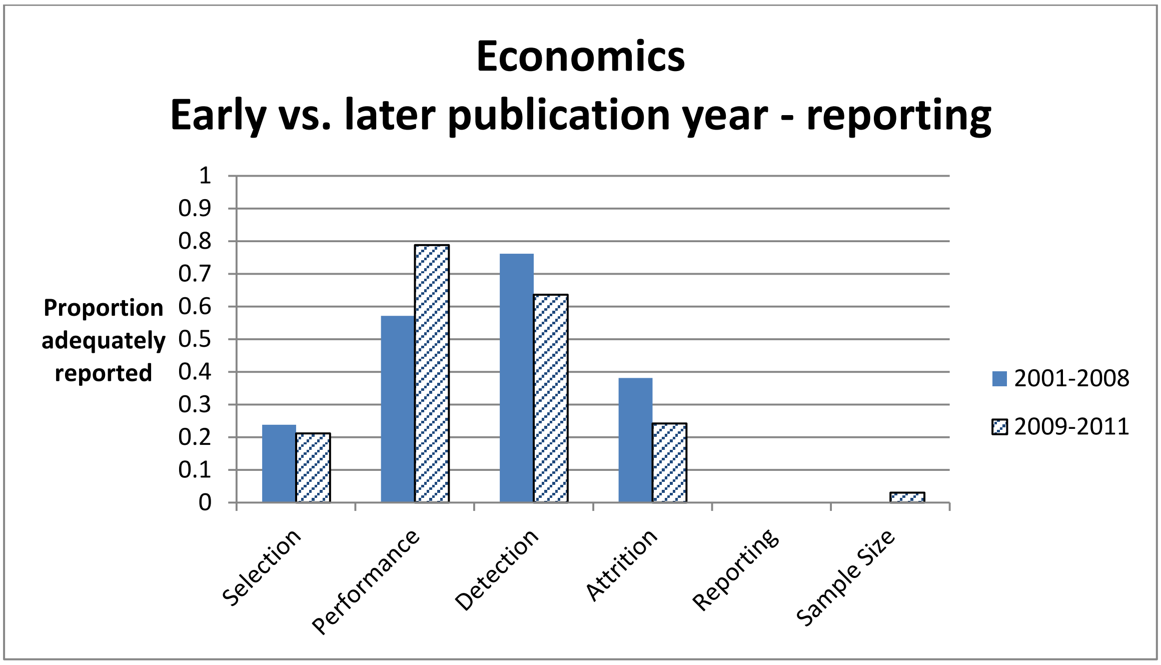

Sample size
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Proportion of papers
Adequately reported
At low risk of bias
Economics
Medicine

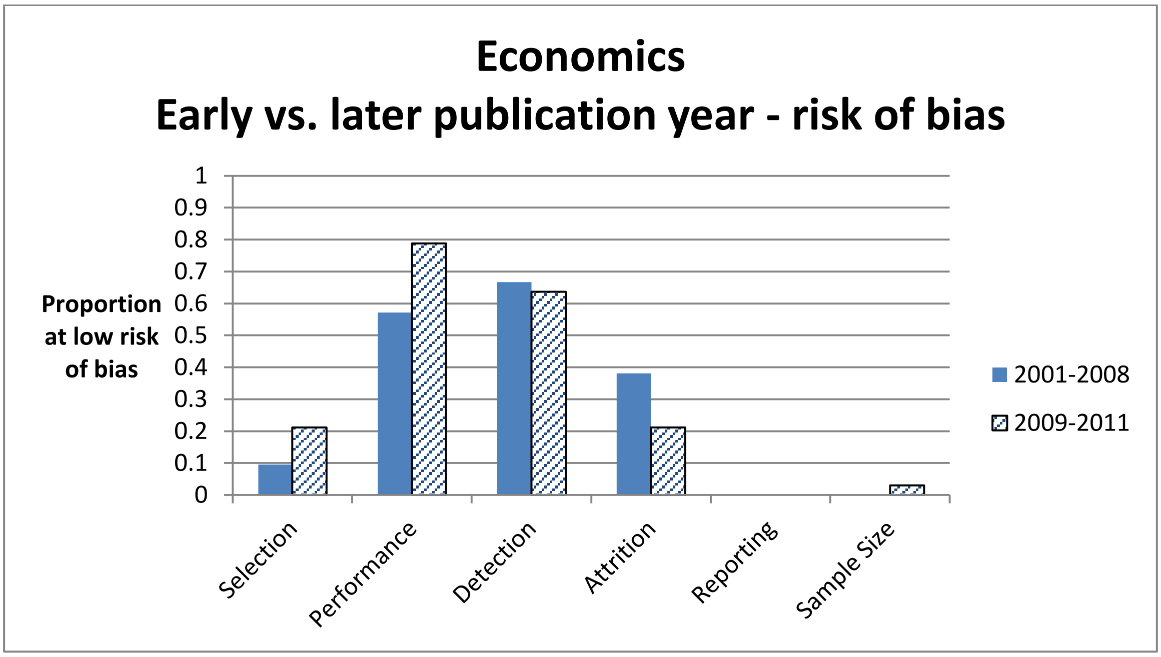
Appendix 3.3: Within-economics subgroup comparisons

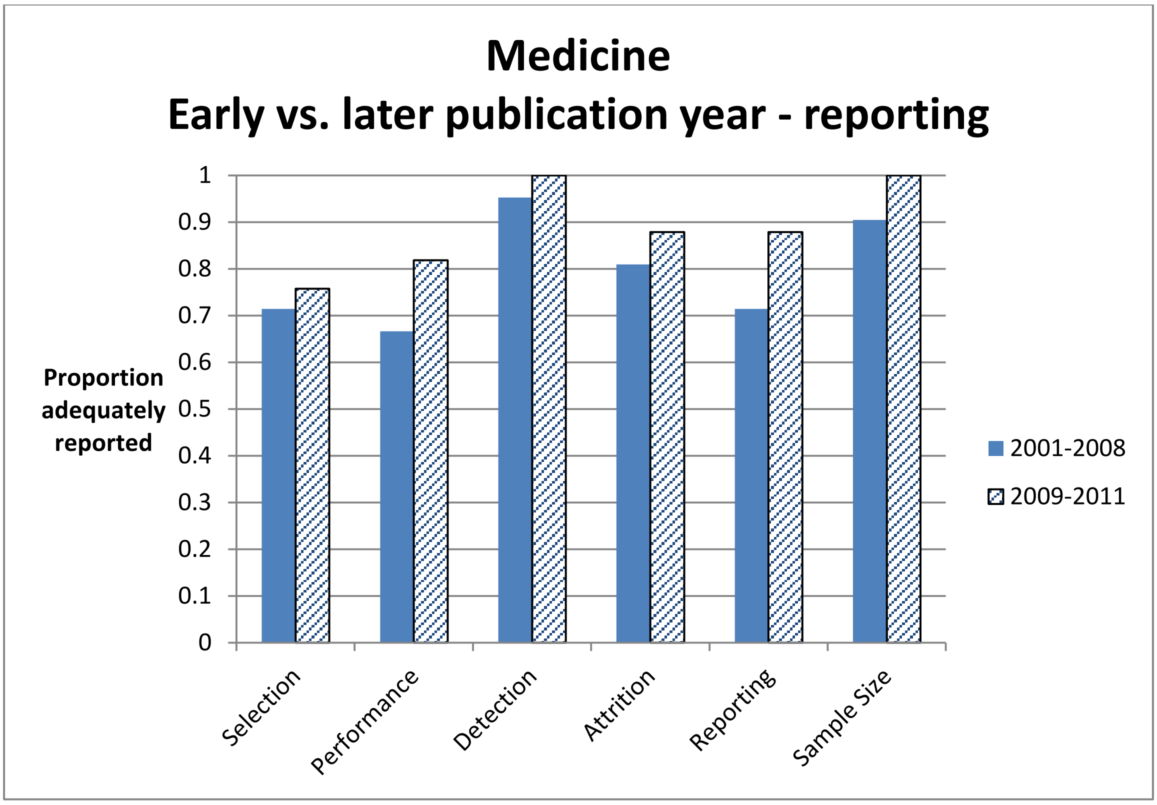


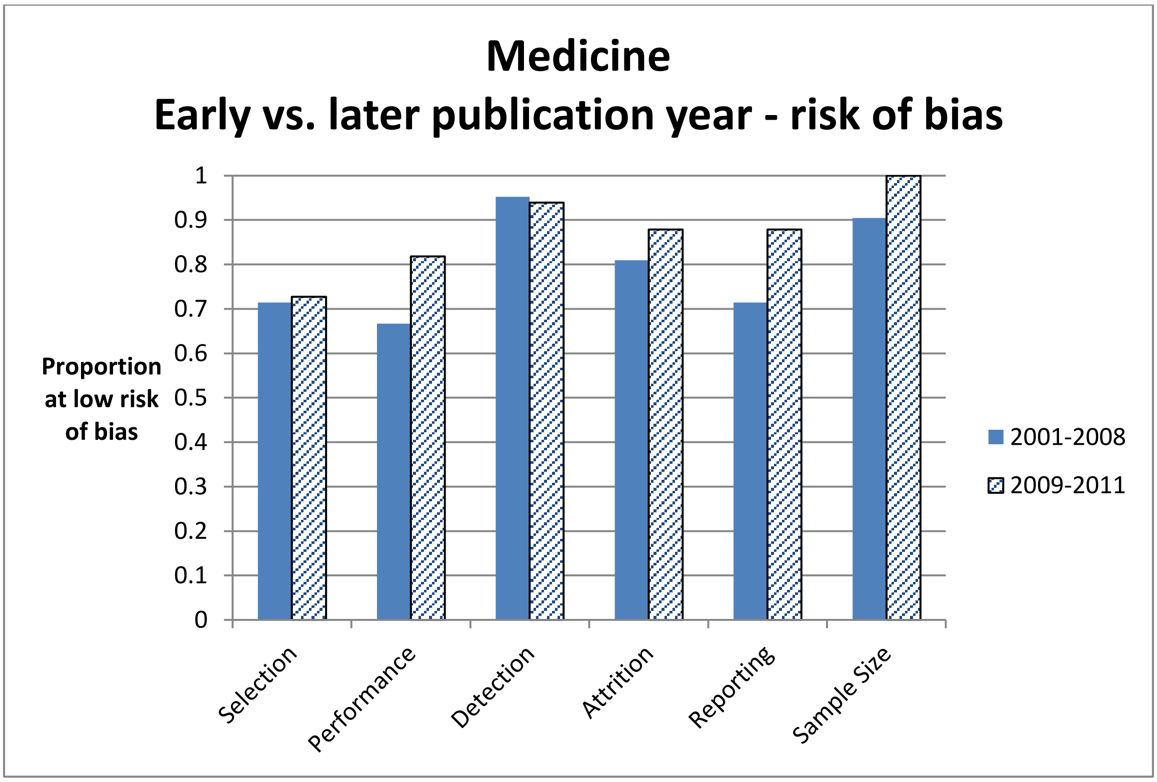


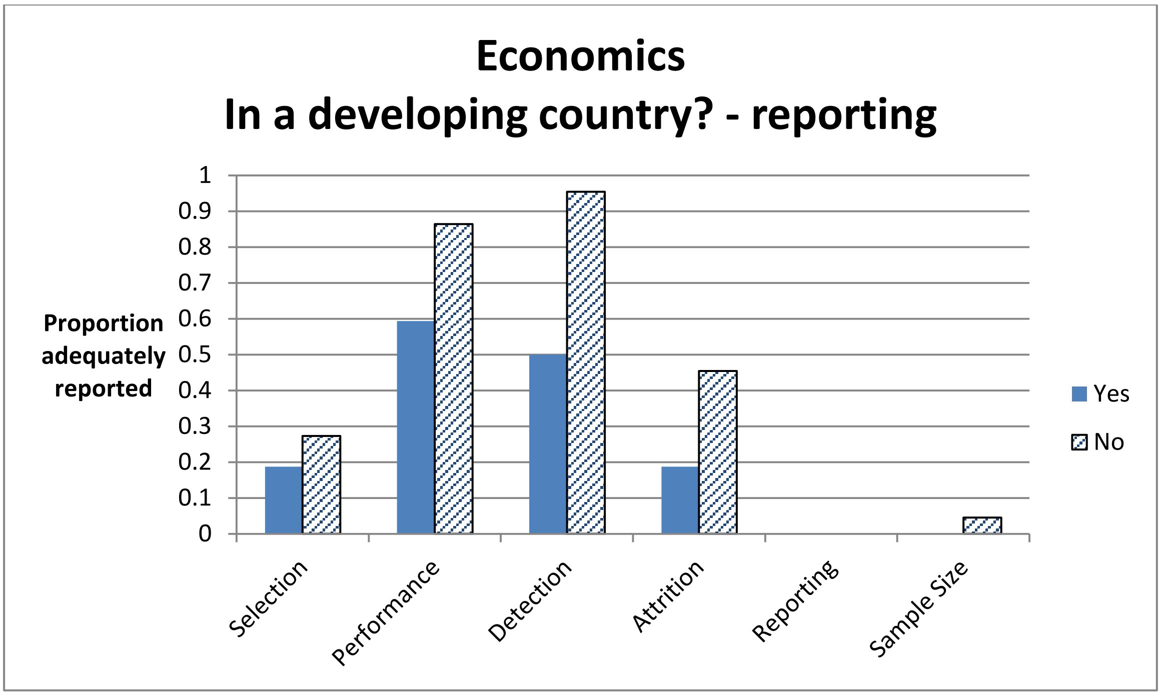
Appendix 3.4: Subgroup comparisons across the two disciplines

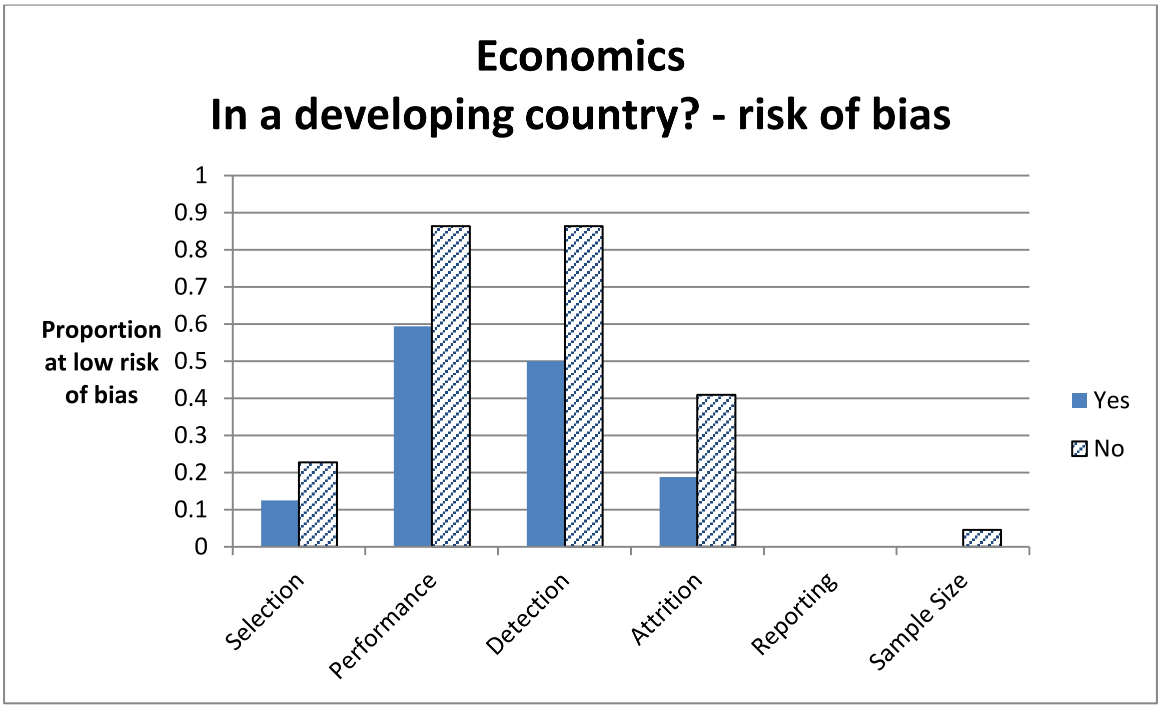


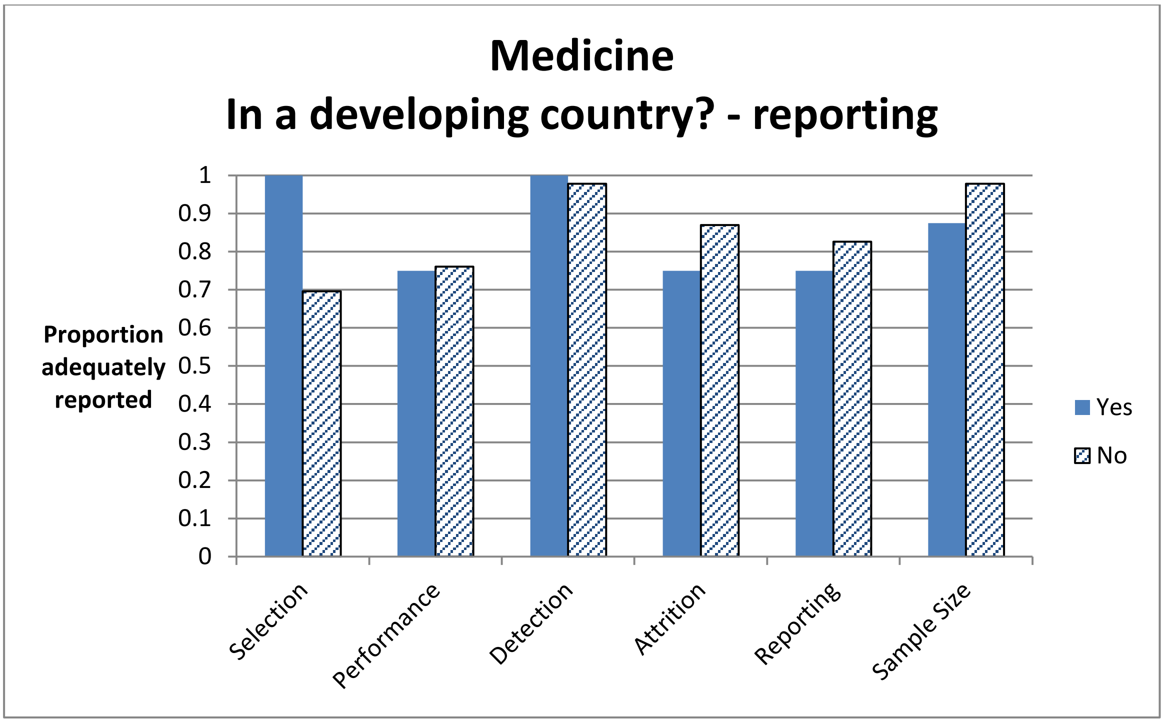


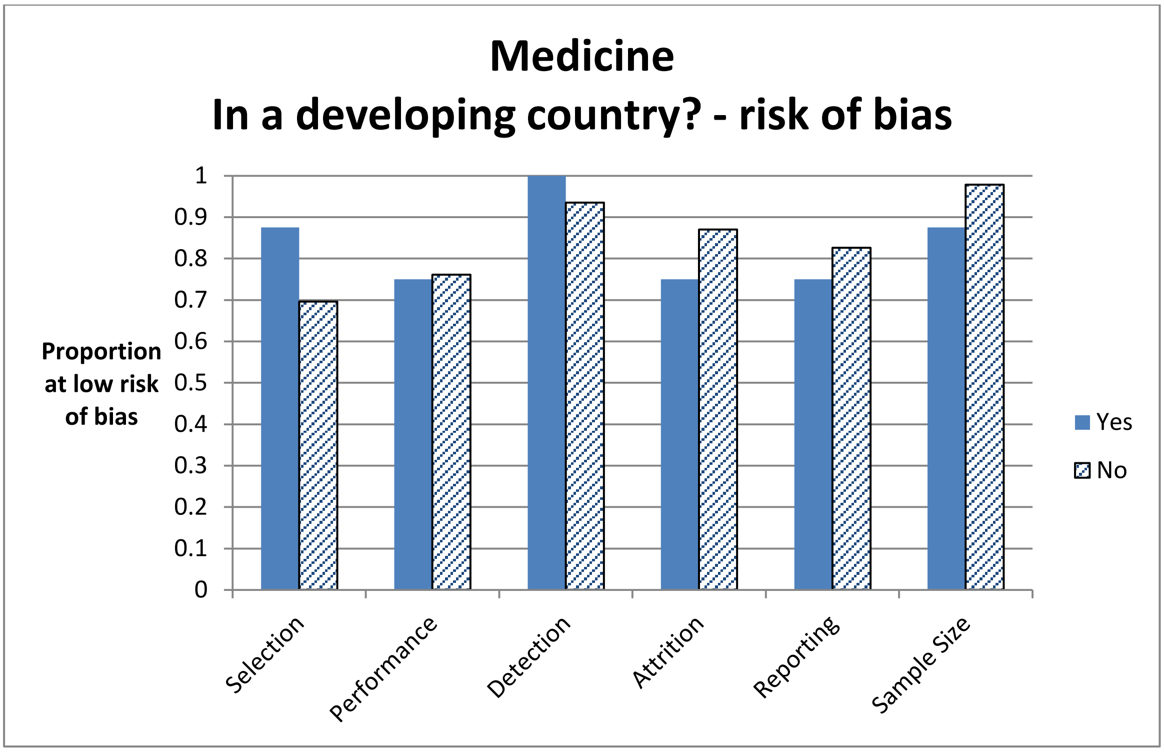

Medicine
Early vs. later publication year - reporting
Proportion adequately reported
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Selection
Performance
Detection
Attrition
Reporting
Sample Size
2001-2008
2009-2011


Medicine
Early vs. later publication year - risk of bias
Proportion at low risk of bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Selection
Performance
Detection
Attrition
Reporting
Sample Size
2001-2008
2009-2011


Economics
In a developing country? - reporting
Proportion adequately reported
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Selection
Performance
Detection
Attrition
Reporting
Sample Size
Yes
No


Economics
In a developing country? - risk of bias
Proportion at low risk of bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Selection
Performance
Detection
Attrition
Reporting
Sample Size
Yes
No


Medicine
In a developing country? - reporting
Proportion adequately reported
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Selection
Performance
Detection
Attrition
Reporting
Sample Size
Yes
No


Medicine
In a developing country? - risk of bias
Proportion at low risk of bias
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Selection
Performance
Detection
Attrition
Reporting
Sample Size
Yes
No

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1239	Richard Layard Dan Chisholm Vikram Patel Shekhar Saxena	Mental Illness and Unhappiness
1238	Laura Jaitman Stephen Machin	Crime and Immigration: New Evidence from England and Wales
1237	Ross Levine Yona Rubinstein	Smart and Illicit: Who Becomes an Entrepreneur and Does it Pay?
1236	Jan-Emmanuel De Neve Ed Diener Louis Tay Cody Xuereb	The Objective Benefits of Subjective Well-Being
1235	Pascal Michaillat Emmanuel Saez	A Model of Aggregate Demand and Unemployment
1234	Jerónimo Carballo, Gianmarco I.P. Ottaviano Christian Volpe Martincus	The Buyer Margins of Firms' Exports
1233	Daniel Fujiwara	A General Method for Valuing Non-Market Goods Using Wellbeing Data: Three-Stage Wellbeing Valuation
1232	Holger Breinlich Gianmarco I. P. Ottaviano Jonathan R. W. Temple	Regional Growth and Regional Decline
1231	Michalis Drouvelis Nattavudh Powdthavee	Are Happier People Less Judgmental of Other People's Selfish Behaviors? Laboratory Evidence from Trust and Gift Exchange Games
1230	Dan Anderberg Helmut Rainer Jonathan Wadsworth Tanya Wilson	Unemployment and Domestic Violence: Theory and Evidence
1229	Hannes Schwandt	Unmet Aspirations as an Explanation for the Age U-Shape in Human Wellbeing
1228	Bénédicte Apouey Andrew E. Clark	Winning Big But Feeling No Better? The Effect of Lottery Prizes on Physical and Mental Health
1227	Alex Gyani Roz Shafran Richard Layard David M Clark	Enhancing Recovery Rates: Lessons from Year One of the English Improving Access to Psychological Therapies Programme

1226	Stephen Gibbons Sandra McNally	The Effects of Resources Across School Phases: A Summary of Recent Evidence
1225	Cornelius A. Rietveld David Cesarini Daniel J. Benjamin Philipp D. Koellinger Jan-Emmanuel De Neve Henning Tiemeier Magnus Johannesson Patrik K.E. Magnusson Nancy L. Pedersen Robert F. Krueger Meike Bartels	Molecular Genetics and Subjective Well-Being
1224	Peter Arcidiacono Esteban Aucejo Patrick Coate V. Joseph Hotz	Affirmative Action and University Fit: Evidence from Proposition 209
1223	Peter Arcidiacono Esteban Aucejo V. Joseph Hotz	University Differences in the Graduation of Minorities in STEM Fields: Evidence from California
1222	Paul Dolan Robert Metcalfe	Neighbors, Knowledge, and Nuggets: Two Natural Field Experiments on the Role of Incentives on Energy Conservation
1221	Andy Feng Georg Graetz	A Question of Degree: The Effects of Degree Class on Labor Market Outcomes
1220	Esteban Aucejo	Explaining Cross-Racial Differences in the Educational Gender Gap
1219	Peter Arcidiacono Esteban Aucejo Andrew Hussey Kenneth Spenner	Racial Segregation Patterns in Selective Universities
1218	Silvana Tenreyro Gregory Thwaites	Pushing On a String: US Monetary Policy is Less Powerful in Recessions
1217	Gianluca Benigno Luca Fornaro	The Financial Resource Curse
1216	Daron Acemoglu Ufuk Akcigit Nicholas Bloom William R. Kerr	Innovation, Reallocation and Growth

The Centre for Economic Performance Publications Unit
Tel 020 7955 7673 Fax 020 7404 0612
Email info@cep.lse.ac.uk Web site http://cep.lse.ac.uk