OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 81, 4 (2019) 0305–9049
doi: 10.1111/obes.12289

Specification Searching and Significance Inflation Across Time, Methods and Disciplines*

Eva Vivalt†

†*Research School of Economics, Australian National University, Acton, ACT 2601, Australia (e-mail: eva.vivalt@anu.edu.au)*

Abstract

This paper examines how significance inflation has varied across time, methods and disciplines. Leveraging a unique data set of impact evaluations on 20 kinds of development programmes, I find that results from randomized controlled trials exhibit less significance inflation than results from studies using other methods. Further, randomized controlled trials have exhibited less significance inflation over time, but quasi-experimental studies have not. There is no robust difference between results from researchers affiliated with economics departments and those from researchers affiliated with other predominantly health-related departments. Overall, the biases found appear much smaller than those previously observed in other social sciences.

I. Introduction

Specification searching is a concern for all quantitative disciplines. However, it is not clear when it is likely to happen. The term 'specification searching' could be used to refer to several phenomena; here, I narrowly consider what I will call 'significance inflation', i.e. any process, such as running multiple regressions and disproportionately reporting those that are significant or collecting more data until results are significant, that leads the statistical significance of reported results to be inflated. This is also known as 'p-hacking'. This paper exploits a new database of published articles and unpublished working papers relating to international development to explore the issue. The data, collected in the process of conducting 20 meta-analyses of development programmes, allow me to test for differences in significance inflation across time, methods and disciplines. I find that

*I am very grateful to the editor, Jonathan Temple, and three anonymous referees for useful comments and suggestions. I also thank Edward Miguel, Bill Easterly, David Card, Ernesto Dal Bó, Hunt Allcott, Elizabeth Tipton, Vinci Chow, Willa Friedman, Xing Huang, Michaela Pagel, Steven Pennings, Edson Severnini, seminar participants at the University of California, Berkeley, Columbia University, New York University, the World Bank, Princeton University, the University of Toronto, the London School of Economics, Cornell University, the University of Ottawa, the Stockholm School of Economics, and participants at the 2015 ASSA meeting and 2013 Association for Public Policy Analysis and Management Fall Research Conference for feedback on an earlier draft of this paper. I am also grateful for the work put in by many at AidGrade to create the data set used in this paper, including Bobbie Macdonald, Diana Stanescu, Cesar Augusto Lopez, Jennifer Ambrose, Naomi Crowther, Timothy Catlett, Joohee Kim, Gautam Bastian, Christine Shen, Taha Jalil, Risa Santoso and Catherine Razeto.

randomized controlled trials (RCTs) exhibit less significance inflation than papers using a quasi-experimental approach. I also find that randomized controlled trials previously exhibited significance inflation, but this decreased over time; in contrast, biases in quasi-experimental studies have, if anything, grown over time. I compare results from studies by economists with results from studies by non-economists, but the differences are insignificant in most cases. In the data set considered, these 'non-economists' are mostly health researchers. Both the economics and non-economics results appear to suffer much less significance inflation than has been found in other social sciences (Gerber and Malhotra, 2008a,b). I consider both published and unpublished papers and find more bias in published papers.

Specification searching has long been seen to be a problem in medicine (e.g. Simes, 1986; Begg and Berlin, 1988) and psychology (Bastardi, Uhlmann and Ross, 2011; Simmons and Simonsohn, 2011). Leamer recognized the problem in economics quite early (Leamer, 1978), and recently there has been new interest in the social sciences (Franco, Malhotra and Simonovits, 2014), including political science (Gerber and Malhotra, 2008a), sociology (Gerber and Malhotra, 2008b), and economics (Brodeur *et al.*, 2016; Bruns, 2017; Ioannidis, Stanley and Doucouliagos, 2017). However, the possibility that there may be differences in significance inflation by method and discipline has not yet fully been explored.

Significance inflation could vary by method and discipline for many reasons. If the journals of different disciplines were to have different selection functions and authors engage in specification searching to try to meet the journal's requirements for publication, this would be sufficient to generate differences in significance inflation. The journals of some fields, for example, may simply be more competitive, raising the bar such that only papers with significant results are published, assuming that significant results are always preferred to insignificant results. Alternatively, it could be the case that some disciplines place more weight on methods and are more likely to accept any well-done RCT; then we would expect RCTs to exhibit less significance inflation in that discipline, as it would be easier for such papers to reach the threshold for publication without significant results. We may also think that the level of significance inflation needed to be competitive at a journal may depend on whether others submitting to the same journals are engaging in it, so different journals or disciplines could be at different equilibria. There are also reasons to believe RCTs suffer less significance inflation independent of journals. Authors that conduct RCTs may be more likely to register a pre-analysis plan, which would serve to diminish the opportunity for running many regressions and selectively reporting results or increasing one's sample size until results are significant. Further, randomization should, in principle, lead to covariate balance across the treatment and control groups; given that one way in which p-hacking may occur is by authors including different combinations of control variables until finding a significant result, randomization could decrease the risk of significance inflation if researchers find it harder to justify including control variables in an RCT.

All these considerations provide reason to suspect significance inflation could differ systematically by method or field, though I will not be able to determine precisely which of these or other factors are responsible for the patterns of specification searching I observe. I also detect more signs of bias in the published literature compared to the unpublished

literature. Specification searching is intimately related to publication bias, as researchers may engage in specification searching in anticipation of journals selectively accepting papers with significant results, but I remain agnostic as to the cause of the observed results.

II. Data

This paper leverages a database of impact evaluation results collected by AidGrade, a US non-profit organization that focuses on gathering the results of impact evaluations and analysing the data, including through meta-analysis (AidGrade, 2018). Its data on impact evaluation results were collected in the course of its meta-analyses from 2012 to 2014.

In the following subsections, I describe the process through which the data were gathered. This is important in determining whether there was likely to have been any selection bias. The process can be thought of as having several stages: first, interventions were selected; then, papers were identified within those interventions; finally, data were extracted from those papers.

It should be noted that the three stages described here (selection of interventions, paper identification and data extraction) are only part of the overall process AidGrade followed to conduct meta-analysis. The latter stages relating to data analysis are not relevant to this paper, but are described elsewhere (Vivalt, 2017). AidGrade followed a unique process for conducting meta-analysis in that it was very inclusive in identifying papers and extracting data, as its goal was to extract results from all papers on a given intervention and to only later categorize and screen the results extracted for meta-analysis; the typical meta-analysis process, in contrast, conducts a much more targeted search and screening of papers to address a narrow set of outcomes. This means that the data used in this paper were relatively minimally screened, as will be described. A companion paper focusing on estimating the heterogeneity in treatment effects within intervention-outcome combinations uses a much smaller and more intensively-screened subset of these data (Vivalt, 2017); in that paper, it was important to ensure the outcome variables used in different studies were precisely comparable, which led most of the data used in this paper to be discarded. This paper is relevant to the project of describing heterogeneity in treatment effects, however, because some of the observed heterogeneity could theoretically be due to specification searching. This paper's results are thus useful in interpreting the results of the companion paper.

Two slightly different approaches were followed for the 10 meta-analyses started in 2012 and 2013. The summary below describes the process for those meta-analyses begun in 2013; corresponding processes for those begun in 2012 are noted where they differ. The selection of interventions was the only stage of the process that differed between the two rounds of meta-analysis.

Selection of interventions

Four AidGrade staff members each independently made a preliminary list of interventions for examination and these lists were then combined; in 2012, there were no staff members and the preliminary list was made solely by the author. Next, pilot searches were done for each topic using SciVerse and Google Scholar to determine if there were likely to be enough impact evaluations for a meta-analysis. As these were not intended to be comprehensive

searches, a low threshold was set of two papers for an intervention to not be rejected at this stage; a more comprehensive search was conducted at a later stage. In 2012, 12 potential interventions were identified by the pilot searches; in 2013, 42 potential interventions.

In 2013, the shortlisted interventions were posted on the AidGrade website and members of the general public were asked to vote on the interventions they wanted covered, in connection with a crowdfunding campaign. The voting window was eight days. Respondents were allowed to select up to three interventions from among the 42 on the short list, with a space provided for adding the 'other' option. 158 individuals cast 452 votes in the timeframe, with 20 selecting the 'other' option.

In 2012, a public vote was also held, but in practice it did not affect the interventions selected; it transpired that lack of overlap on common outcome variables would prohibit subsequent meta-analysis of two of the 12 interventions, and this exactly determined the 10 interventions ultimately selected from the list of 12, as will be described. How outcome variables were defined must first be discussed. Three different variables were created to label the outcomes at varying degrees of specificity. First, there was a set of 'strict' outcomes, which measured the exact same thing across papers. For example, height in centimeters. There was also a set of 'loose' outcomes, which captured outcomes that could be defined slightly differently across different studies. For example, one study might consider participants to have anaemia if their haemoglobin was less than some threshold X; another study might consider participants to have anaemia if their haemoglobin was less than some different threshold Y. These outcomes are clearly related, but results for anaemia could differ between two studies simply because of how the measure was defined in each study. Finally, there was a set of 'broad' outcomes that merely captured whether the outcome could be thought of as an 'economic', 'educational', or 'health' outcome.[1] At the 'strict' outcome level, there was very little overlap across studies; as outcomes were defined more broadly, there was more overlap across studies. For meta-analysis, it was important to compare similar outcomes, so a criterion was set that, after the search and screening stages, relevant papers would be scanned for prospective future 'strict' outcomes held in common and if at least three papers covering a common outcome variable were not found that intervention would not be included and another intervention would be selected; in 2012, this determined the 10 interventions selected from the shortlist of 12.

In 2013, the shortlist (prescreening) contained 42 interventions but, as in 2012, there was only capacity to cover 10 interventions, so it was decided to partially randomize. The randomization was done to ensure as much balance as possible between those topics included and excluded; however, the most popular topic from the public vote, women's empowerment programmes, was automatically selected and left out of the randomization process, which matched interventions using nearest neighbour matching prior to randomization.[2]

[1] Not all outcomes fell into one of these categories. For example, marriage and other social outcomes were not considered to fit in any of these broad outcome categories, even though they might indirectly affect economic, educational or health outcomes.

[2] To obtain balance among the interventions included and excluded, each shortlisted topic was matched with another of the shortlisted topics based on how many likely impact evaluations the pilot searches identified for each; how many votes they received in the public vote; the overall theme of the interventions (e.g. education, health) according to the database of an external organization, AidData, after matching the interventions to AidData activity codes; and the recent aid commitments for the intervention as reported in AidData's database. The theme had to match exactly within each pair. For each of the three other factors, each topic was assigned a score on an index between zero and

TABLE 1

List of development programmes covered

2012	2013
Conditional cash transfers	Contract teachers
Deworming	Financial literacy training
Improved stoves	HIV education
Insecticide-treated bed nets	Irrigation
Microfinance	Micro health insurance
Safe water storage	Micronutrient supplementation
Scholarships	Mobile phone-based reminders
School meals	Performance pay
Unconditional cash transfers	Rural electrification
Water treatment	Women's empowerment programmes

Notes: This table lists the development programmes considered in this paper. Three titles here may be misleading. 'Mobile phone-based reminders' refers specifically to SMS or voice reminders for health-related outcomes. 'Women's empowerment programmes' required an educational component to be included in the intervention and it could not be an unrelated intervention that merely disaggregated outcomes by gender. Finally, 'micronutrient supplementation' was initially too loosely defined; this was narrowed down to focus on those providing zinc to children, but the other micronutrient papers are still included in the data used in this paper.

After matching, half the interventions (one from each pair) were randomly discarded to shorten the shortlist. Since women's empowerment programmes were guaranteed to be included as an intervention, nine additional interventions had to be selected from the remaining list. Having found that the interventions selected in 2012 had few outcome variables in common, these nine interventions were selected to be those on the list that were covered by the most studies in the pilot searches. Nonetheless, some of these interventions were covered by as few as three impact evaluations.

The topics that were ultimately selected for study in each round are listed in Table 1.

Identification of papers

A comprehensive literature search was then done using a mix of the search aggregators SciVerse, Google Scholar, and EBSCO/PubMed. The online databases of the Abdul Latif Jameel Poverty Action Lab, Innovations for Poverty Action, the Center for Effective Global Action and the International Initiative for Impact Evaluation were also searched for completeness. Finally, the references of any existing systematic reviews or meta-analyses were collected.

one representing where it stood among the other interventions; the index took the value: (topic value – minimum value among topics)/(maximum value among topics – minimum value among topics). 32 topics were successfully matched in this way using nearest neighbour matching without replacement. The remaining unmatched topics were singletons under their respective themes. For example, if there were an odd number of health-related interventions, the last health-related intervention would be by itself after others were matched. These last topics were independently randomized.

Any impact evaluation which appeared to be on the intervention in question was included, barring those in developed countries.[3] Both published papers and working papers were included. As mentioned, the search and screening criteria were deliberately broad. The full text of the search terms and inclusion criteria for all 20 topics are available as an appendix. Screening was done in several stages: first, the titles were screened; then, for any papers passing the title check, the abstracts were screened; finally, for those papers passing the title and abstract check, AidGrade staff looked for the full text of the paper, and the full text of those papers that were found was screened.

This process resulted in a list of studies predominantly authored by researchers in economics-related disciplines. The other main discipline represented in the data was health. To examine field-specific biases, coders were instructed to determine whether a majority of each paper's authors were formally affiliated with an economics or economics-related institution, such as a department of agricultural economics. Those that did not are considered here as 'non-economics': they consist almost exclusively of researchers affiliated with schools of public health or medicine.

Data extraction

All data were entered independently by two different coders and any discrepancies were reconciled by a third. Coders followed a convention to extract those results with the fewest control variables. It was thought that this might minimize bias due to specification searching, since one easy way for researchers to engage in specification searching is by including additional controls. Further, where results were presented separately for multiple subgroups, coders were similarly advised to err on the side of caution and to collect both the aggregate results and results by subgroup, except where the author appeared to be including a subgroup only because results were significant within that subgroup.[4] We might expect that these two conventions excluded some cases of specification searching; if so, the results presented here should be considered as lower bounds for the *extent* of specification searching. It is not immediately clear whether this might affect observed *differences* between disciplines and methodologies. However, it should be noted that quasi-experimental and economics papers tended to have specifications with more control variables excluded due to this coding rule, so using those results with the fewest controls might serve to mask differences between methods and disciplines if we think including more control variables helps to enable significance inflation.

This data extraction process resulted in multiple results being extracted from each paper. For example, a study could present results for multiple treatment arms, different subgroups, different specifications, or different time periods. When conducting a meta-analysis, decisions have to be made about which results to include, and in the companion paper I followed a protocol to identify or construct a single result per intervention-outcome-paper combi-

[3] The World Bank's country classification system was used for this, with 'high-income' countries excluded (World Bank, 2015).

[4] For example, if an author reported results for children aged 8–15 and then also presented results for children aged 12–13, only the aggregate results would be recorded, but if the author presented results for children aged 8–9, 10–11, 12–13, and 14–15, all subgroups would be coded as well as the aggregate result when presented. Authors only very rarely reported isolated subgroups, so this was not a major issue in practice.

nation (Vivalt, 2017). However, many more results were extracted than ultimately used for meta-analysis, including results for all subgroups and all time periods. When results for several different treatment arms were reported, results for each arm were collected, even if these results were based on the same control group, though the fact that they shared the same control group would be captured in another variable. For papers that reported results that were obtained through different methods, such as a paper reporting both results from an RCT and a quasi-experimental differences-in-differences approach, results were collected from the approach considered 'more rigorous' according to an internal hierarchy.[5] If any doubt remained on which of two results to extract, coders were instructed to extract both and flag them.

Similarly, when treatments were presented crossed with other treatments or interacted with other variables, these interaction coefficients and standard errors were still extracted, even though they would be difficult to interpret in a meta-analysis. However, the focus of the data collection was meta-analysis of the effects of certain interventions, so if a paper dealt *only* with interaction terms or spillovers its results were not included. This could plausibly make it less likely to find any evidence of significance inflation, as it means that data were sometimes gathered for results that may not have been the primary focus of that paper's authors, and it seems reasonable to believe that authors might face more pressure to obtain significant results for their primary findings.

Results could be reported in several ways. Rather than reporting a coefficient and standard error from a regression, a paper might directly report a z-statistic or a number of other variables, such as the treatment and control group mean, the treatment and control group standard deviation, and the number of observations in each of the treatment and control group. Coders were instructed to collect the data as it was reported to minimize coding errors, using a data extraction form, and these data were later converted to z-statistics.

For the sake of this paper, it is important to note that while most results were reported in a way that immediately made their significance clear, for some results, significance was not immediately conveyed and had to be calculated. If the significance could be calculated without making any assumptions, it was, such as dividing a regression coefficient by its standard error. However, for an important subset of results, significance could only be calculated by making an additional assumption. These results were those that were initially reported in tables providing the mean, standard deviation, and number of observations in each of the treatment and control groups. For this way of reporting results, which was quite common in the non-economics literature, the treatment effect could be calculated as the difference between the treatment and control group means, but a pooled standard error had to be calculated from the standard deviation and the number of observations in the treatment and control groups. In calculating this, I assume the effects in the treatment and control groups were independent, allowing this statistic to be calculated in a straightforward way.[6] Independence is a reasonable assumption for RCTs and the vast majority of papers

[5] In order from those considered most to least rigorous: RCT, regression discontinuity design, differences-in-differences, matching.

[6] $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where s_1 is the estimate of the standard deviation of the treatment group, n_1 is the number of observations in the treatment group, and likewise s_2 and n_2 represent these values for the control group.

TABLE 2

Descriptive statistics

Category	No. results	No. papers	Number of results per paper		
			25th percentile	50th percentile	75th percentile
Main sample					
All papers	8,738	473	3	8	20
RCTs	6,978	335	4	9	24
Non-RCTs	1,760	138	2	6	15
Economics	5,499	240	3	8	25
Non-economics	3,239	233	4	9	18
Published	6,434	366	3	8	18
Unpublished	2,304	107	3	8	27
Extended sample					
All papers	11,970	584	4	9	24
RCTs	10,015	439	4	11	29
Non-RCTs	1,955	145	3	6	15
Economics	5,690	244	3	8	25
Non-economics	6,280	340	5	10	24
Published	9,645	475	4	10	24
Unpublished	2,325	109	3	8	25

Notes: This table restricts attention to those results for which full information about significance is available ('main' sample) or can be imputed with an assumption of independence ('extended' sample).

reporting results in this format were from non-economics RCTs. While it would be more suspect for non-RCTs, only 195 results from 13 non-RCTs were reported in this way.[7] Still, to guard against the possibility that this subset of results is affecting my findings, these results were not included in the main tables in this paper, dropping 3,232 results from 184 papers.[8] Instead, I report a full set of alternative tables in the appendix that include those results whose significance was imputed as described above.[9]

In summary, information was gathered in a highly inclusive way. From the 635 papers from which data were extracted (474 RCTs, 161 non-RCTs, 261 economics, 374 non-economics), 17 were dropped for not being an impact evaluation of the effect of the intervention itself, only presenting results in which the intervention was interacted with some other factor or only presenting results in which the intervention was compared to another treatment (13 RCTs, four non-RCTs, five economics, 12 non-economics). Then 34 were dropped for not presenting enough information about significance for P-values to be calculated to at least three digits (22 RCTs, 12 non-RCTs, 12 economics, 22 non-economics). Instead, some papers merely noted whether a result was significant or not at a given level of significance (e.g. $P < 0.05$), without providing information that could be used

[7] The other results came from 171 RCTs. Only 12 economics papers reported at least some results in this way, along with 172 non-economics papers.

[8] This does not mean all results from those papers were necessarily dropped; some papers reported results in different ways for different outcome variables.

[9] The most exact significance information was used wherever possible so that, for example, if a paper reported the treatment and control group mean, standard deviation, and number of observations, but also reported the exact P-value in the text, the P-value would be used and the result included in the main tables.

to determine the exact *P*-value, provided sufficient information from which to determine significance levels only for significant results, provided two-digit *P*-values, or occasionally presented results without any information at all about significance. Further, as discussed, the significance of 3,232 results could only be calculated with error. In total, after these papers were discarded, 8,738 results with full information about significance levels remained from 473 papers, or 11,970 from 584 papers using the 'extended' sample that includes the values imputed with error. Table 2 provides details on the number of results per study in the final sample across each of the main subgroups of interest.

III. Method

This paper examines specification searching by comparing the number of barely significant results with the number of barely insignificant results around the conventional cut-off significance level of 5%. I will argue that if one looks at the distribution of *z*-statistics in a body of literature, one should expect to see roughly comparable numbers of results just on either side of any given threshold when restricting attention to a narrow enough band centred on that threshold. The paper will consider the ranges 2.5%, 5%, 10%, 15% and 20% above and below $z = 1.96$, in turn, and examine whether results follow a binomial distribution around 1.96 as we would expect in the absence of bias. For example, the 2.5% range would run from 1.911 to 2.009. This is subsequently referred to as a caliper test. This basic approach was introduced by Gerber and Malhotra (2008a,b).

When using caliper tests, one should carefully consider the issues arising from having multiple coefficients coming from the same papers. In particular, a handful of papers could theoretically drive results. Gerber and Malhotra (2008a,b) address the issue by breaking down their results by the number of coefficients contributed by each paper, so as to separately show the results for those papers that contribute one coefficient, two coefficients, and so on. I instead aggregate the results by paper in robustness checks, so that, for example, a paper with four coefficients below the threshold and three above it would be counted as 'below'.[10] I use this approach as it can mitigate the risk that a few papers are responsible for much of the effect, but it does discard information and results in fewer observations. A potential alternative would be to weight observations by the inverse of the number of results taken from that paper, however, due to some cells containing a relatively small number of observations, I prefer to use exact tests, which are incompatible with weighting. Further, there is an argument to be made for considering results for multiple outcome variables from the same paper equally, since each of these outcome variables may contribute to the literature independently of how many other outcome variables are considered by the same paper, and the primary reason a paper reports multiple results in my data is that these results represent effects on different outcome variables. In a sense, it is interesting both whether the typical result suffers from significance inflation and whether the typical paper suffers from significance inflation. I thus report both types of results, with results based on the disaggregated data presented as the main results.

[10] Ties are excluded.

TABLE 3

Caliper tests: by result

Caliper	*(1) Over*	*(2) Under*	*(3) Prop.*	*Caliper*	*(4) Over*	*(5) Under*	*(6) Prop.*	*(7) Diff.*
RCTs				*Non-RCTs*				
2.5%	103	65	0.61***	2.5%	36	10	0.78***	−0.17**
5%	149	139	0.52	5%	61	28	0.69***	−0.17***
10%	280	310	0.47	10%	87	65	0.57*	−0.10**
15%	414	462	0.47	15%	124	100	0.55	−0.08**
20%	536	615	0.47**	20%	159	140	0.53	−0.07**
Economics				*Non-economics*				
2.5%	97	52	0.65***	2.5%	42	23	0.65**	0.00
5%	151	118	0.56*	5%	59	49	0.55	0.02
10%	255	264	0.49	10%	112	111	0.50	−0.01
15%	360	396	0.48	15%	178	166	0.52	−0.04
20%	454	518	0.47**	20%	241	237	0.50	−0.04

Notes: This table shows the number of results that fall into each caliper over time, as well as the proportion of results that fall above the threshold for significance within the caliper and the difference of this proportion across RCTs and non-RCTs. Stars indicate whether the proportions are statistically significantly different from 0.5 (columns 3 and 6) and from each other (column 7). ***Significant at $P < 0.01$, **significant at $P < 0.05$, *significant at $P < 0.1$.

IV. Results

The first results are presented in Table 3. Quasi-experimental studies, which will be referred to as 'non-RCTs', appear to suffer from significance inflation. RCTs also show signs of significance inflation in the narrowest caliper, but in general perform much better in terms of the magnitudes of the share of results just over the threshold and their significance. The difference between RCTs and non-RCTs in the share of results just over the threshold for significance is formally tested in the last column using Fisher's exact test. It should be recalled that since the distribution of the z-statistics is skewed, we should expect to see fewer results just over as opposed to just under the threshold for significance as the caliper size increases, which is indeed the case.

There are no significant differences between results from papers written by authors affiliated with economics departments and those written by others. However, almost all quasi-experimental studies were conducted by economists. This suggests caution in interpreting results. Table A1 in the appendix presents results disaggregated by both discipline and method used. Among economics papers, RCTs have significantly fewer results above the threshold for significance than non-RCTs in all calipers; among non-RCTs, the sample size is small, but there is some suggestion that economics papers have more results above the threshold for significance than non-economics papers. However, among RCTs, non-economics papers have more results above the threshold for significance in the 15% and 20% caliper.

Turning to the time dimension, Tables 4 and 5 show the percent of results that were just above, as opposed to just below, the threshold for significance within various time periods. Again, in the absence of bias, one would expect 50% of results to fall on either side; perhaps slightly less within the wider calipers, due to the natural slope of results when the

TABLE 4

Caliper tests: by result over time, RCTs vs. non-RCTs

RCTs				*Non-RCTs*				
Caliper	*(1) Over*	*(2) Under*	*(3) Prop.*	*Caliper*	*(4) Over*	*(5) Under*	*(6) Prop.*	*(7) Diff.*
1990–99								
2.5%	9	4	0.69	2.5%	0	0	N/A	N/A
5%	12	7	0.63	5%	0	0	N/A	N/A
10%	25	25	0.50	10%	0	0	N/A	N/A
15%	41	35	0.54	15%	0	0	N/A	N/A
20%	52	50	0.51	20%	0	0	N/A	N/A
2000–09								
2.5%	63	27	0.70***	2.5%	31	8	0.79***	−0.09
5%	82	73	0.53	5%	45	22	0.67***	−0.14*
10%	152	166	0.48	10%	64	53	0.55	−0.07
15%	223	260	0.46	15%	90	79	0.53	−0.07
20%	303	338	0.47	20%	119	112	0.52	−0.04
2010+								
2.5%	31	34	0.48	2.5%	5	2	0.71	−0.24
5%	55	59	0.48	5%	16	6	0.73*	−0.24**
10%	103	119	0.46	10%	23	12	0.66*	−0.19**
15%	150	167	0.47	15%	34	21	0.62	−0.14*
20%	181	227	0.44**	20%	40	28	0.59	−0.14**

Notes: This table shows the number of results that fall into each caliper over time, as well as the proportion of results that fall above the threshold for significance within the caliper and the difference of this proportion across RCTs and non-RCTs. Stars indicate whether the proportions are statistically significantly different from 0.5 (columns 3 and 6) and from each other (column 7). ***Significant at $P < 0.01$, **significant at $P < 0.05$, *significant at $P < 0.1$.

null hypothesis is true. There is not enough information to say anything about non-RCTs and economics papers in the 1990s using these data, but RCTs and non-economics papers can still be compared across the three time periods, with the caveat that the composition of RCTs by discipline clearly changes across the time periods.

These tables show that in recent years, specification searching seems to have decreased for RCTs in the narrowest caliper where it was previously significant; this decrease is significant when explicitly tested (Table A2). In contrast, non-RCTs may have even exhibited greater biases in recent years, especially in larger calipers, though the changes among non-RCTs over time are not significant. While there are few results in the smallest calipers, the magnitudes observed in the larger calipers would be consistent with a story in which non-RCTs were treated with increased skepticism over time, especially if they reported marginally significant results. For example, if it became widely perceived that a z-statistic of 1.97 was not credible, a natural response by researchers engaging in specification searching would be to report fewer of these values and more z-statistics like $z = 2.20$. These results are based on relatively few observations, so they are not definitive. The difference between RCTs and non-RCTs in 2010-14 is statistically significant at $P < 0.10$ or lower for the 5%, 10%, 15% and 20% calipers, however, when RCTs and non-RCTs were mostly not significantly different despite the larger sample sizes in 2000–09 (Table 4).

TABLE 5

Caliper tests: by result over time, economics vs. non-economics

Economics				*Non-economics*				
Caliper	*(1) Over*	*(2) Under*	*(3) Prop.*	*Caliper*	*(4) Over*	*(5) Under*	*(6) Prop.*	*(7) Diff.*
1990–99								
2.5%	0	0	N/A	2.5%	9	4	0.69	N/A
5%	0	0	N/A	5%	12	7	0.63	N/A
10%	0	0	N/A	10%	25	25	0.50	N/A
15%	0	0	N/A	15%	41	35	0.54	N/A
20%	0	0	N/A	20%	52	50	0.51	N/A
2000–09								
2.5%	66	20	0.77***	2.5%	28	15	0.65*	0.12
5%	90	60	0.60**	5%	37	35	0.51	0.09
10%	147	149	0.50	10%	69	70	0.50	0.00
15%	201	227	0.47	15%	112	112	0.50	−0.03
20%	264	293	0.47	20%	158	157	0.50	−0.03
2010+								
2.5%	31	32	0.49	2.5%	5	4	0.56	−0.06
5%	61	58	0.51	5%	10	7	0.59	−0.08
10%	108	115	0.48	10%	18	16	0.53	−0.05
15%	159	169	0.48	15%	25	19	0.57	−0.08
20%	190	225	0.46*	20%	31	30	0.51	−0.05

Notes: This table shows the number of results of studies by authors in economics and non-economics departments that fall into each caliper over time, as well as the proportion of results that fall above the threshold for significance within the caliper and the difference of this proportion across economics and non-economics papers. Stars indicate whether the proportions are statistically significantly different from 0.5 (columns 3 and 6) and from each other (column 7). ***Significant at $P < 0.01$, **significant at $P < 0.05$, *significant at $P < 0.1$.

No differences between economics and non-economics papers are significant (Table 5), though 77% of results in economics papers in the 2.5% caliper in 2000–09 fell over the threshold, significantly different from 50% at $P < 0.001$. These results may be driven by the results for non-RCTs, as previously suggested. From 2010 on, the non-economics papers had a larger magnitude of results above the threshold for significance, although insignificantly.

Overall, these findings mark a great departure from Gerber and Malhotra's findings regarding the political science (2008a) or sociology (2008b) literature. A direct comparison cannot be made, as the data collection methods were different, but there is a striking difference between these results and their rows of proportions statistically different from 0.5 with $P < 0.001$.

Analyses collapsing results by paper and disaggregating by publication status are included in the appendix (Tables A3–A4). The estimates of the proportion of studies above the threshold appear slightly smaller when collapsing by paper, with the share above the threshold decreasing in most cases. Results are also less significant. Nonetheless, the results still show the expected decline in the share of results over the threshold as the caliper

widens, and they still show RCTs have fewer results just over as opposed to just under the threshold for significance than non-RCTs, though this is not significant. One change when collapsing results by paper is that non-economics papers appear to have a relatively larger share of results above the threshold than economics papers, though this is again generally insignificant. This suggests caution when interpreting the earlier results which showed, if anything, economics papers exhibiting more signs of bias; that finding could be the result of a few authors.

Published results appear more biased than unpublished results (Table A4). Non-RCTs and RCTs still are significantly different among unpublished papers in the 2.5% ($P < 0.05$), 5% ($P < 0.001$), and 10% calipers ($P < 0.1$), and among published papers in the 5% ($P < 0.1$), 15% ($P < 0.1$), and 20% calipers ($P < 0.05$), with RCTs having fewer results above the threshold than non-RCTs. There are not many unpublished non-economics papers; among published papers, economics papers have more results above the threshold, but this is insignificant. A limitation is that this study cannot speak to papers that were not only not published but also 'file drawered', i.e. not even available as a working paper.

V. Discussion

Four important issues should be further discussed: why caliper tests are justified; differences by how results were reported; the possibility of selection bias due to how studies and their results were selected for inclusion; and the relatively low power of some of the tests.

The use of caliper tests requires some justification, as the validity of these tests was not thoroughly discussed in the seminal papers introducing them (Gerber and Malhotra, 2008a,b). We will first consider the case in which, for every hypothesis tested in the data, the null hypothesis is true. We will then extend this argument to deal with the more plausible scenario in which some null hypotheses are false.

If the null hypothesis is always true, it is quickly clear that the z-statistics should be equally distributed around a given threshold for a small enough band centred on that threshold. The z-statistics would be normally distributed in large samples, and the probability density function of that distribution would be smooth.

In the case that the null hypothesis is sometimes false, it is not clear what shape the distribution of z-statistics will take.

I will put forward three arguments that even in the case in which the null hypothesis is sometimes false, we should not expect to see *the distribution of results observed* in the absence of specification searching. In particular, the distribution I observe shows more results above than below the threshold of $z = 1.96$, and not elsewhere, which I take as evidence of bias towards significant results. Second, in the absence of bias, it is unclear why RCTs do not typically show a jump in marginally significant results compared to marginally insignificant results but quasi-experimental studies do. Third, it does not appear to be the case that studies are being powered just enough so as to yield significant results, which could also theoretically lead to a jump in the number of marginally significant results compared to marginally insignificant results. The subsequent paragraphs expand on each point.

First, if the z-statistics exhibit a jump at $z = 1.96$ but are otherwise smoothly decreasing at nearby values, it is hard to see what could explain that other than some kind of bias

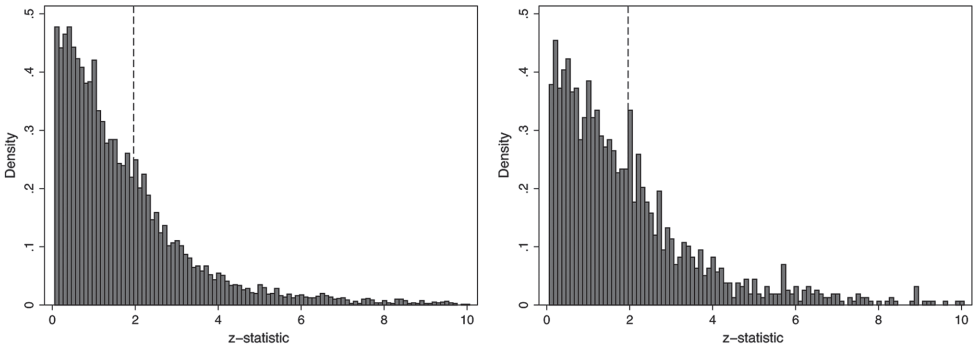

Figure 1. Quasi-experimental studies show more signs of bias
Notes: This figure plots the distribution of z-statistics in the data for all studies considered in this paper (left) and non-RCTs (right). A dashed line is drawn at $z = 1.96$ in each plot, and each bar represents a 0.1 range of z-statistics, starting at 0.06 so as to be able to clearly distinguish the threshold of $z = 1.96$. It should be noted that though I find that quasi-experimental studies exhibit signs of specification searching, they still seem to suffer much less bias than the results reported in Gerber and Malhotra (2008a).

towards significant results. In particular, while we might imagine that the distribution of z-statistics when the null is false could take any number of shapes, we may be willing to make the assumption that this distribution will be smooth. If the distribution is smooth, the difference in the probability that an observed z-statistic x falls in an interval of width ε just below z and the probability that x falls in the adjacent interval of the same width just above z will approach zero as ε approaches zero. Even if we are unwilling to assume the distribution is smooth, we may be willing to assume that the distribution will not exhibit a distinct jump at precisely $z = 1.96$ and nowhere else in the absence of bias. If we observe there is a different pattern of results just where we would expect there to be under bias, it would be reasonable to take that as evidence of bias.

Figure 1 plots the distribution of z-statistics that I observe in the data. Figure 1a shows the distribution of z-statistics when including results from all papers; Figure 1b shows the distribution when including results only from those papers that were not RCTs. Figure 1a shows a fairly smooth distribution, as we might expect to observe in the absence of bias. There is perhaps a small bump in the distribution around $z = 1.96$, though it is not as prominent as in other work. In Figure 1b, however, there is a visible jump in the z-statistics right at $z = 1.96$ that is greater than at any other point in the distribution.

Second, if there is an alternative story as to why the z-statistics take the distribution they do, that story should explain the features of the observed data. *A priori*, we might expect that RCTs would exhibit fewer traces of bias due to their being more likely to be published independent of their results and due to their increased rigor perhaps making specification searching more difficult. If some other factor is driving a spike in z-statistics just above $z = 1.96$, this factor would have to differently affect RCTs and non-RCTs. This would seem to weigh against the story in which the pattern is a function of some null hypotheses being false, unless one believed that quasi-experimental studies were more likely to have false null hypotheses.

Finally, one might be concerned that selection bias is driving the observed distribution of z-statistics rather than specification searching. Namely, it is possible that researchers are selecting to conduct studies when they believe the effects will be marginally significant.

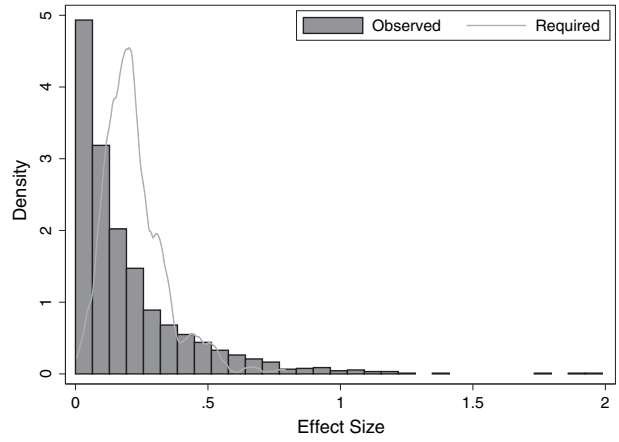

Figure 2. Observed vs. required effect sizes
Notes: This figure plots the distribution of effect sizes found among the subset of the data which could be standardized, resulting in a smaller data set of 1,405 observations. A kernel density function is overlaid using an epanechnikov kernel with bandwidth 0.0213, showing the distribution of effect sizes that researchers would have needed in order to obtain a power of 0.8 given the sample sizes selected, assuming they had control over sample sizes. Effect sizes larger than two are not shown for either distribution for legibility.

This is the strongest competing hypothesis, as it could in theory explain a jump in the density of z-statistics just above $z = 1.96$ as well as why results for RCTs and non-RCTs might exhibit different patterns.

On the one hand, even if the results are driven by selection bias rather than specification searching, the trajectory and distribution of this bias across disciplines and time periods would still be interesting and worthy of study. However, there are many reasons to believe the specific jump in z-statistics at $z = 1.96$ is not being driven by selection bias.

First, the studies in the data set appear underpowered on net. If researchers were selecting projects based on their ability to obtain significant results, the studies ought to be better-powered. While I do not know the power calculations that researchers might have made *a priori*, for a subset of the data I can graph the distribution of effect sizes that researchers would have needed for their studies to have had a power of 0.8 for a two-sided test with $\alpha = 0.05$, given the sample sizes observed. I can then overlay the distribution of effect sizes actually found (Figure 2). The two distributions look quite different and suggest the majority of results are underpowered.

This is not conclusive in itself, as the majority of results could be underpowered while researchers still select on expected significance in a subset of cases. For example, a study could be powered to detect effects on one set of outcome variables but also report results for other outcome variables for which the study does not have adequate power. However, there are several assumptions underlying the selection bias story that are unlikely to hold. First, researchers would have to have control over the sample size, effect size, or the standard deviation of the outcome variable. Sample size is often constrained by funding considerations and not something researchers have fine control over; researchers are even less likely to have precise control over the effect size or standard deviation of the outcome variable. Second, in order to cause a distinct spike at $z = 1.96$ and nowhere else, researchers

would have to have a very clear idea of the effect of the study, within a narrow range of values, before it was implemented.

In other data, researchers have been shown to have inaccurate priors as to the effects of various programmes, generally over-estimating the effect sizes found. Groh *et al.* (2012) survey 136 attendees at research seminars – two at academic institutions and two in international organizations – as well as readers of the World Bank's Development Impact Blog and find that after describing the intervention and setting but before presenting results, the median guess of each of six treatment effects differs from the true treatment effect by a minimum of approximately 70%. DellaVigna and Pope (2018) also ask 314 experts from behavioural conferences to predict the effects of various behavioural treatments on the effort exerted by MTurk participants. They provide the experts with the results from three benchmark treatments to help them calibrate how responsive participants were to past incentives and then ask them to predict the effort participants exert in 15 other treatments. Perhaps due to some combination of their providing sample past experimental results and the fact they are estimating behavioural responses, which may have less variance than the treatment effects of typical development interventions, the average absolute error in individual forecasts of treatment effects is only 8%. However, if one were to test for a difference between each experimental treatment and the first, basic treatment, even this small difference in forecasted treatment effect would translate to a forecasted z-statistic that differed from the true z-statistic by an average of 4.1. The smallest magnitude in the error in forecasted z-statistics among the 15 experimental treatments would be 2.2.

It thus does not seem very plausible for researchers to be able to guess the effects of the programme so accurately as to select a sample that would result in a z-statistic between 1.96 and 2.009 and so fall within the 2.5% band but above the significance threshold. For the median result's sample size of 1,041 observations, this range of z-statistics corresponds to an effect size between 0.0607 and 0.0623. I cannot completely rule out selection, but if manipulation is occurring, it would seem easier for it to occur via specification searching than through guessing an effect size to this degree of accuracy and precision.

Turning to discuss selection bias, AidGrade's data set was compiled to conduct meta-analyses, and therefore the papers and results included in it were not selected with this study in mind. However, the selection process could still theoretically bias this paper's results. In particular, the interventions considered were selected from a larger list of possible interventions, partially on the basis of that intervention being covered by a relatively large number of papers (although that 'large number' was just three). Therefore, the interventions selected may be those that were particularly easy to implement or popular as social programmes; those that were expected to find large effects; or those that had particularly contentious or heterogeneous results leading to increased researcher incentives to continue to study them.

It is unclear how the ease of implementation or popularity of a social programme could affect significance inflation. If an intervention were expected to have large effects on a set of outcomes, that might obviate the need for specification searching to obtain significant results as it would shift the distribution of P-values upwards for those outcomes. However, we might expect most outcomes studied by papers on that intervention to be unaffected, as it seems unlikely that even the best programmes would find a large effect along all

outcomes studied. A related concern is that we might expect somewhat more specification searching among interventions with contentious results.

To further explore the issue, I present a set of results broken down by whether the intervention-outcome combination was covered by a relatively 'large' (>10) or 'small' ($\leqslant 10$) number of papers and whether it was one of the 'earlier' (first 10) or 'later' papers on that topic, to look for suggestive evidence on the sign and magnitude of this potential bias.[11] I assign these variables within intervention–outcome combination, since we might expect that even among papers covering the same intervention there is scope for papers to differentiate themselves, so that it would not matter much whether other papers had been written on that intervention and what they had found. For example, if there were already a lot of papers on the effects of conditional cash transfer programmes on children's educational outcomes, there could still be scope for additional papers on the health effects of conditional cash transfer programmes.

Recall that three different types of outcomes were assigned, at varying levels of specificity: 'strict', 'loose', and 'broad' outcomes. I use the 'broad' outcomes for two reasons. First, they may better capture the novelty of a result. For example, we might think that if there were already a lot of papers on the effect of a particular intervention on school enrolment rates (a 'strict' outcome), a paper on the effect of that intervention on school attendance rates (a different 'strict' outcome) would not really be novel but it would be counted as such by the 'strict' or 'loose' outcomes. If I used the 'broad' outcome labels, both these outcomes would be classified as 'educational' outcomes, and the last study would rightly be considered not so novel. As a second benefit, this approach allows me to leverage many more observations, given that many more observations were assigned a 'broad' outcome than a 'strict' or 'loose' one, due to the fact that 'strict' and 'loose' outcome labels were only assigned when multiple papers within an intervention covered the same outcome variable and there was not much overlap across studies.

Results are presented in Table A5. Results from papers on topics covered by more papers appear to exhibit more bias in the smallest caliper; there are also more results above the threshold among early papers in the smallest caliper, though insignificantly more results above the threshold among later results in the largest caliper. The greatest magnitude of these differences is in the smallest caliper; differences in other calipers appear small.

There are many factors that could be driving these results. For example, if we believed the difference between early and late results in the smallest caliper was real, that could be consistent with the theory of a 'decline effect', whereby early, promising results were later overturned; it would also be consistent with a story in which the earliest papers on a subject are written by academic researchers under pressure to obtain good publications and later papers are written by researchers without such career concerns. These results suggest caution in interpreting some of the other results in this paper: if results from intervention–outcome combinations covered by a large and small number of papers were different, we

[11] It is possible that if estimates of specification searching were being magnified by the inclusion of some interventions with contentious results, the contention would have resolved itself after the first few papers and hence the threshold of 10 papers might be too large. However, I pick this number since the median number of studies on a given intervention-outcome is itself quite large, at 34, so effects found for smaller numbers would be less likely to be driving results and the sample would be smaller if a lower threshold were used.

might worry that the extent of specification searching is overstated, as the selection process prioritized interventions covered by a relatively large number of papers.

However, there is not much difference in the number of papers covering a topic across the dimensions considered in this paper. For the sake of reference, the median result among all results collected from economics papers was on an intervention-outcome combination covered by 31 papers, just lower than the median for non-economics papers of 34; the median non-RCT result is also on an intervention-outcome covered by 31 papers, compared to a median for RCTs of 34. Nonetheless, the possibility remains that results could be different on a sample that considered more seldom-studied interventions. My results thus remain tentative, but the fact that outside the smallest caliper there are not large differences is somewhat reassuring, given some of the strongest results comparing RCTs with non-RCTs showed significant differences in the larger calipers.

Next, we may wish to consider the manner in which results were reported. In particular, recall that while the main tables excluded those results whose significance had to be imputed with error from the mean, standard deviation and number of observations in each of the treatment and control groups, this was a fairly substantial number of results, so tables in the appendix repeat the analysis including these results (Tables A6–A13). Interestingly, when including these cases, the proportion of results over the threshold for significance falls in almost all cases and results often become less significant, though economics papers become significantly more biased than non-economics papers in the smallest caliper in 2000–09. The reduced significance could reflect error from assuming independence, since assuming independence would result in larger standard errors than would be true if the dependence were known and taken into consideration. However, there are other potential causes. This trend would also be consistent with a story in which specification searching became more likely where the significance of a result was immediately apparent. It is also possible that authors chose how they reported results depending on their significance. For example, providing a regression result and standard errors with stars might emphasize a result's significance, while providing the mean, standard deviation and number of observations separately for the treatment and control group could be used to make a result's insignificance less obvious. While I cannot determine which of these competing hypotheses is driving the observation that these results appear less significant, that observation is interesting and suggests room for further research to determine whether mode of reporting might causally affect significance inflation.

Finally, a note should be made about the power of the tests presented in this paper. If the power of the tests varies across method, discipline or time, this could call into question the cross-method, cross-discipline or cross-time comparisons. While post hoc power calculations cannot be used to interpret existing results (Hoenig and Heisey, 2001), for a hypothetical future study to be powered at 0.8 to detect a difference between 65% of one sample of results falling above the threshold for significance and 50% of another sample of results falling above the threshold for significance, each sample should contain about 181 results assuming $\alpha = 0.05$ and a two-sided test; assuming $\alpha = 0.1$, 142 results.[12] It should be noted that Fisher's exact test, which these calculations are based on, is a conservative test. If we were merely comparing one sample with 65% of results above the

[12] This is calculated by simulation with 10,000 runs for each value.

threshold for significance to a hypothetical 50% proportion in that caliper, a total of 85 observations would be needed for the same power for a two-sided test with $\alpha = 0.05$, or 67 for a two-sided test with $\alpha = 0.1$.

Overall, while some cells considered in this paper contain few observations, in many cases there are sufficient observations that the cross-method, cross-discipline and cross-time differences are significant. The low power of some comparisons remains a limitation of this paper.

VI. Conclusions

This paper finds that studies using randomized experiments exhibit less specification searching than those that do not. However, these biases appear less pronounced than has previously been found in some of the other social sciences, and there appears to be little difference between papers written by researchers in economics-related disciplines and papers written by researchers in other fields like public health. The data include results from both published and unpublished papers, and results from unpublished papers show less evidence of bias than results from published papers.

A second contribution is that specification searching is shown to not be static, but a bias that evolves. In particular, RCTs have exhibited significantly less bias over time, while quasi-experimental studies have, if anything, exhibited more pronounced biases over time. There are a few possible intuitive explanations for these results. First, it could be the case that standards are becoming relatively higher for RCTs than for papers using quasi-experimental methods, which are perhaps increasingly published in lower-ranked journals and facing less scrutiny or attention. Alternatively, quasi-experimental studies may be facing more pressure to find strongly significant results in order to be taken seriously. Both possibilities point to the importance of researcher incentives, which should be taken into consideration to address the problem.

References

AidGrade (2013). AidGrade process description.

AidGrade (2018). AidGrade impact evaluation data, version 1.5.

Bastardi, A., Uhlmann, E. L. and Ross, L. (2011). 'Wishful thinking: belief, desire, and the motivated evaluation of scientific evidence', *Psychological Science*, Vol. 22, pp. 731–732.

Begg, C. and Berlin, J. (1988). 'Publication bias: a problem in interpreting medical data', *Journal of the Royal Statistical Society. Series A*, Vol. 151, pp. 419–463.

Brodeur, A., Le, M., Sangnier, M. and Zylberberg, Y. (2016). 'Star wars: the empirics strike back', *American Economic Journal: Applied Economics*, Vol. 8, pp. 1–32.

Bruns, S. (2017). 'Meta-regression models and observational research', *Oxford Bulletin of Economics and Statistics*, Vol. 79, pp. 637–653.

DellaVigna, S. and Pope, D. (2018). 'Predicting experimental results: who knows what?', *Journal of Political Economy*, Vol. 126, pp. 2410–2456.

Franco, A., Malhotra, N. and Simonovits, G. (2014). 'Publication bias in the social sciences: unlocking the file drawer', *Science*, Vol. 345, pp. 1502–1505.

Gerber, A. and Malhotra, N. (2008a). 'Do statistical reporting standards affect what is published? Publication bias in two leading political science journals', *Quarterly Journal of Political Science*, Vol 3, pp. 313–326.

Gerber, A. and Malhotra, N. (2008b). 'Publication bias in empirical sociological research: do arbitrary significance levels distort published results?', *Sociological Methods & Research*, Vol. 37, pp. 3–30.
Groh, M., Krishnan, N., McKenzie, D. and Vishwanath, T. (2012). *Softskills or Hard Cash? The Impact of Training and Wage Subsidy on Female Youth Employment in Jordan*, Policy Research Working Paper No. 6141.
Hoenig, J. M. and Heisey, D. M. (2001). 'The abuse of power: the pervasive fallacy of power calculations for data analysis', *The American Statistician*, Vol. 55, pp. 1–6.
Ioannidis, J., Stanley, T. D. and Doucouliagos, H. (2017). 'The power of bias in economics research', *The Economic Journal*, Vol. 127, pp. F236–F265.
Leamer, E. (1978). *Specification Searches: Ad hoc Inference with Nonexperimental Data*, John Wiley & Sons, Inc., New York, NY.
Simes, J. (1986). 'Publication bias: the case for an international registry of clinical trials', *American Society of Clinical Oncology*, Vol. 4, pp. 1529 –1541.
Simmons, J. and Simonsohn, U. (2011). 'False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological Science*, Vol. 22, pp. 1359–1366.
Vivalt, E. (2017). *How Much Can We Generalize from Impact Evaluations?* Working Paper. Unpublished manuscript.
World Bank (2015). *Country and Lending Groups*. Retrieved (September 1, 2015), available at http://data.worldbank.org/about/country-and-lending-groups.

Supporting Information

Additional supporting information may be found in the online version of this article:

Appendix A: Additional results.
Appendix B: Excerpt from AidGrade (2013).
Appendix C: The search terms and inclusion criteria for each topic.
Appendix D: Bibliography of included papers.
Appendix E: The coding manual.