The Trajectory of Specification Searching and Publication Bias Across Methods and Disciplines

Eva Vivalt

Australian National University

February 12, 2018

Abstract

This paper examines how specification searching and publication bias have varied across time, disciplines, and methods. Studies by researchers affiliated with economics departments appear to exhibit more bias than those affiliated with schools of public health or medicine, but less than has been observed for other social sciences. Most of this gap appears related to the methods used. Randomized controlled trials, in particular, exhibit less bias than studies using other methods. Increased skepticism of results from quasi-experimental studies may paradoxically have led to more extreme specification searching or publication bias among them over time.

1 Introduction

Specification searching and publication bias are concerns for all quantitative disciplines. However, it is not clear when they are likely to happen. Researchers working in different academic fields might face different pressures leading to distinct patterns of biases. This paper considers several types of studies and finds that specification searching and publication bias vary by time period, methodology and discipline.

Specification searching and publication bias have long been seen to be a problem in medicine (*e.g.* Begg and Berlin, 1988; Simes, 1986) and psychology (Simmons and Simonsohn, 2011; Bastardi, Uhlmann and Ross, 2011). There has been growing interest as well 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.*, 2012). This paper exploits a database of 705 published articles and unpublished working papers relating to international development. The data provide a long-run view of how these biases might evolve over time due to different pressures. Both randomized controlled trials (RCTs) and quasi-experimental studies are included, and we observe different patterns in each. Further, since both economists and researchers in other fields such as health are represented in this database, often studying the same types of interventions, we can test whether different disciplines appear subject to different biases.

These biases could vary by method and discipline for several reasons. Imagine the following selection process: journals have a preference for publishing significant results (publication bias) and, given that, authors engage in specification searching to try to meet the journal's requirements for publication. There are many possible ways in which both the selection functions of journals and the responses by authors could differ. To give an illustrative example, if the journals of different disciplines were to have different selection functions, this would be sufficient to generate differences in both publication bias and specification searching. Some fields may simply be more competitive, raising the bar such that only good papers with significant results are published. 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 specification searching and publication bias, as it would be easier for such papers to get over the publication threshold without significant results. We may also think that the level of specification searching 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. Authors that conduct RCTs may also be more likely to pre-register a pre-analysis plan, which would also serve to diminish the opportunity for specification searching.

Impact evaluations in economics have recently grown exponentially, both in number and in terms of the resources devoted to them. This adds to the importance of quantifying the biases in them and discovering where they are most likely to appear. While these studies are still growing, a few thousand are already complete, providing a body of work with which to examine to what extent these studies suffer from these kinds of biases.

We find that studies done by researchers affiliated with economics departments show slightly more signs of bias than in health, though less than has been seen in political science or sociology. Further, randomized controlled trials exhibit less bias than studies using other methods. Interestingly, the biases of quasi-experimental studies appear to have grown over time, perhaps in response to increased skepticism of results.

We detect an over-abundance of barely significant results in both unpublished as well as published literature, suggesting the observed patterns are not due to selective publication alone. However, specification searching is intimately related to publication bias, as researchers may engage in specification searching in anticipation of encountering later publication bias, so we cannot cleanly disentangle the two and remain agnostic as to the cause of the observed results.

2 Data

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

To mitigate concerns about selection, the process governing the selection of papers and extraction of results will be briefly discussed. Further information is provided in a set of appendices described in Appendix A.

Two slightly different processes 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 provided where they differ. The selection of interventions was the only stage of the process that substantively differed between the two rounds of meta-analysis. The topics that were ultimately selected for study in each round are listed in Table 1.[1] Four AidGrade staff members each independently made

Table 1: List of Development Programs 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 programs

[1]Three titles here may be misleading. "Mobile phone-based reminders" refers specifically to SMS or voice reminders for health-related outcomes. "Women's empowerment programs" 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.

a preliminary list of interventions for examination; the lists were then combined and pilot searches done for each topic using SciVerse and Google Scholar to determine if there were likely to be enough impact evaluations for a meta-analysis.[2] In 2012, the author identified 12 potential topics by the pilot searches; in 2013, 42 topics were identified.

The shortlisted topics were posted on the AidGrade website and members of the general public were asked to vote on the topics they thought were the most relevant, in connection with a crowdfunding campaign. The voting window was eight days. Respondents were allowed to select up to three topics from among the 42 on the short list, with a space provided for adding an "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 bind since it transpired that lack of overlap on common outcome variables was a constraint. 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 3 papers covering a common outcome variable were not found that topic would not be included; this ultimately determined the 10 interventions selected.

In 2013, there were sufficiently many topics remaining to randomize among the shortlisted topics to obtain the final list, while ensuring as much balance as possible between those topics included and excluded, and also acceding to the vote for the most popular topic, women's empowerment programs.[3] This step was not conducted in 2012.

[2]As these were not intended to be comprehensive searches, a low threshold was set of 2 papers for an intervention to not be rejected at this stage; a more comprehensive search was conducted at a later stage.

[3]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 0 and 1 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 neighbor 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

This process yielded a randomized list of topics to cover, to which the winner of the popular vote, women's empowerment programs, was added. Given capacity constraints, 10 interventions were selected from this list to be covered in 2013. Having found that the interventions covered in 2012 had few outcome variables in common, these 10 interventions were selected to be those on the list that were covered by the most studies in the pilot searches.

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.

Any impact evaluation which appeared to be on the intervention in question was included, barring those in developed countries. Any paper that tried to consider the counterfactual was considered an impact evaluation. Both published papers and working papers were included. The search and screening criteria were deliberately broad. The full text of the search terms and inclusion criteria for all 20 topics are available online (Vivalt, 2016).

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.

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,

randomized.

as one easy way for researchers to engage in specification searching is by including additional controls, while recognizing that there are many other ways in which specification searching may occur. 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 only including a subgroup because results were significant within that subgroup.[4] We might expect that these two conventions exclude some cases of specification searching, in which case the results presented here should be considered as lower bounds for the extent of specification searching.

To help to address the time dimension in this paper, two sets of older papers were added to the AidGrade data. The older papers are those based on two large data sets that were available in the 1980s and 1990s and which had been extensively exploited: the Indonesian Family Life Survey (IFLS) and data from the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT).

Any paper citing one of these two data sets as a data source was considered for inclusion. The other criterion for inclusion was that the results were not simply summarizing the data but attempting to test a hypothesis, bearing in mind that many of the earlier papers strove to test hypotheses simply by running a regression with controls or looking for statistically significant correlates of the variables of interest. Data were extracted following the previously described protocols; Appendix C provides additional detail.

In total, this process resulted in 13,250 results with full information about significance levels across 592 papers.[5] Table 1 provides details on the number of results per study across each of the main subgroups of interest.

[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 rarely reported isolated subgroups, so this was not a major issue in practice.

[5]Some papers merely note whether a result is significant or not at a given level of significance (e.g. $p < 0.05$), without providing information that could be used to determine the exact p-value, and occasionally results are presented without any information at all about significance.

Table 2: Descriptive Statistics

			Number of results per paper		
Category	N results	N papers	25th percentile	50th percentile	75th percentile
All papers	13250	592	4	10	28
RCTs	9958	421	4	12	30
Non-RCTs	3292	171	4	9	24
Economics	6768	253	4	9	32
Non-Economics	6482	339	5	12	27
Published	10875	494	4	10	28
Unpublished	2375	98	4	9	32

This table restricts attention to those results for which full information about significance is available, *i.e.* the set of results that is considered in the rest of this paper.

3 Method

This paper examines specification searching and publication bias by comparing the number of barely significant results with the number of barely insignificant results around the conventional cut-off significance level of 5%. We 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 centered on that threshold. The paper will then 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 approach requires justification. 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 centered on that threshold. If the null hypothesis is always true, the z-statistics are normally distributed, and

the probability density function of the distribution is smooth.

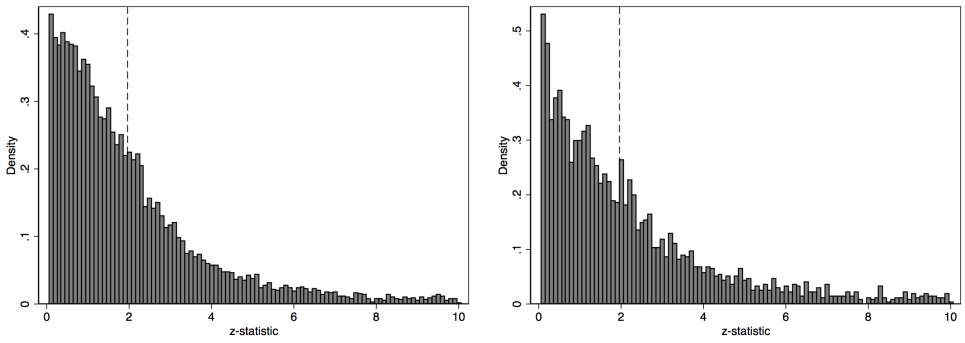
In the case that the null hypothesis is sometimes false, it is not clear what shape the distribution of z-statistics will take.

We 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 that we observe* in the absence of specification searching and publication bias. In particular, the distribution we observe shows more results above than below the threshold of z=1.96, and not elsewhere, which we 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 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 i just below z and the probability that x falls in the adjacent interval of the same width just above z will approach 0 as i approaches 0. 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 we observe in the data. Figure 1a

Figure 1: Recent Quasi-Experimental Studies Show More Signs of Bias



This figure plots the distribution of z-statistics in the data for all studies considered in this paper (left) and recent 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. "Recent" here is defined as starting in 2005, to subdivide the data into two halves chronologically. It should be noted that though we find that quasi-experimental studies exhibit signs of specification searching or publication bias, they still seem to suffer much less bias than the results reported in Gerber and Malhotra (2008a).

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. Figure A.1 in the appendix shows a similar figure from Gerber and Malhotra (2008a), in which this type of spike is even more apparent.

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. 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 for some reason quasi-experimental studies were more likely to have false null hypotheses.

Finally, we may be concerned that rather than specification searching or publication bias driving the observed distribution of z-statistics, it is being caused by selection bias. Namely, it is possible that researchers are selecting to conduct studies when they believe the effects will be marginally significant. This is the strongest competing hypothesis that we can imagine, as it could in theory explain a jump in the density of z-statistics just above z=1.96 and nowhere else 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 or publication bias, the trajectory and distribution of this bias across disciplines and time periods would still be interesting and worthy of study. However, we have many reasons to believe the specific jump in z-statistics at z=1.96 is not being driven by selection bias.

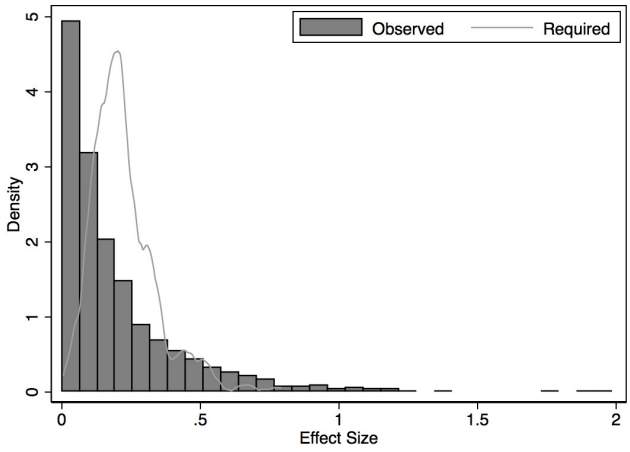
First, the studies in the data set appear underpowered on net. If researchers were selecting projects based on their ability to obtain significant results, the studies ought to be better-powered. While we do not know the power calculations that researchers might have made *a priori*, for a subset of the data we can graph the distribution of effect sizes that researchers would have had to have had in mind 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. We 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 de-

viation 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.

We note that in other data, researchers have been shown to have inaccurate priors as

Figure 2: Observed vs. Required Effect Sizes



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 1,461 observations. A kernel density function is overlaid using an epanechnikov kernel with bandwidth 0.0212, showing the distribution of effect sizes that researchers would have had to have had in mind 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 2 are not shown for either distribution for legibility.

to the effects of various programs, 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 (2016) also ask 314 experts from behavioral conferences to predict the effects of various behavioral 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 behavioral 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 program 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 study's sample size of 553 observations, this range of z-statistics corresponds to an effect size between 0.0833 and 0.0854. We 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.

One final consideration relating to caliper tests should be discussed before turning to the results. When doing the caliper tests used in the rest of the paper, we need to also carefully consider the issues arising from having multiple coefficients coming from the same papers. In particular, we would not want a handful of papers to drive results. Gerber and Malhotra (2008a; 2008b) 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. This paper instead aggregates

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". Ties are excluded. This approach can mitigate the risk that one or two of the papers are responsible for much of the effect, however, it does discard information and results in fewer observations. Results based on the unaggregated data are presented first as the main results.

4 Results

The first results are presented in Table 2. Quasi-experimental studies, which will be referred to as "non-RCTs", appear to suffer from bias, but RCTs perform much better. 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 for a wide enough band, which is indeed what we see for RCTs. These findings, especially for RCTs, 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 the difference between these results and their rows of mostly $p < 0.001$ results, reproduced in the appendix (Table A.1), is striking.

Results from papers written by authors affiliated with economics departments also exhibited more specification searching or publication bias than results from those written by authors from other disciplines (or "non-economic" disciplines). However, almost all quasi-experimental studies were conducted by economists. This suggests caution in interpreting results. When one restricts attention to only RCTs and considers the narrowest calipers, where any observed differences are especially difficult to attribute to anything except specification searching and publication bias, the economics papers still have a slightly greater share of papers just above the threshold for significance than just below it, compared to the non-economics papers (Table A.2). However, results remain only suggestive.

Turning to the time dimension, results here tell an interesting story. Tables 3 and 4

show the percent of results that were just above, as opposed to just below, the threshold for significance, within various bands. Again, in the absence of bias, we would expect 50% to fall on either side; perhaps slightly less within the wider bands, due to the natural slope of results when the null hypothesis is true.

These tables show that in recent years, specification searching and publication bias seem to have stayed roughly the same for RCTs. In the narrowest band, it may have even fallen: 60% of results were just over as opposed to just under the threshold in 2000-2009, compared to 50% from 2010 on, a difference significant in a t-test at $p<0.10$. In contrast, non-RCTs have exhibited greater biases in recent years, especially in larger calipers. In the largest caliper, 51% of results fell just over as opposed to just under the threshold in 2000-2009, compared to 73% from 2010 on, a difference significant at $p<0.001$. It thus appears as if non-RCTs try to hide this bias more or are biased in a different way. For example, it is plausible that if everyone thought a z-statistic of 1.97 was not credible, fewer papers would report these kinds of values, but more would report $z=2.20$. The difference between RCTs and non-RCTs in 2010-2014 is statistically significant at $p<0.05$ or lower for the 10%, 15% and 20% calipers. Results also appear to vary by discipline. The results in economics papers showed more bias in the 2.5% and 5% caliper in 2000-2009, yet more bias in the 20% caliper in 2010-2014. As there are very few non-RCTs in other disciplines, these results seem to be driven by the results for non-RCTs; Table A.2 illustrates.

Analyses collapsing results by paper and disaggregating by publication status are included in the appendix (Tables A.3 - A.4). Results exhibit similar trends when collapsing by paper. The relatively low number of unpublished papers in the database limits the ability to do subgroup analysis, but results for unpublished papers are similar to those of published papers on the aggregate: non-RCTs appear more biased than RCTs. Published papers appear more biased, on the whole, than unpublished papers. 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.

Table 3: Caliper Tests: By Result

	Over Caliper	Under Caliper	p-value
All studies			
2.5% Caliper	181	123	<0.01
5% Caliper	286	252	
10% Caliper	525	530	
15% Caliper	779	795	
20% Caliper	1135	1131	
RCTs			
2.5% Caliper	135	106	<0.10
5% Caliper	205	206	
10% Caliper	401	435	
15% Caliper	590	640	
20% Caliper	821	891	<0.10
Non-RCTs			
2.5% Caliper	46	17	<0.001
5% Caliper	81	46	<0.01
10% Caliper	124	95	<0.10
15% Caliper	189	155	<0.10
20% Caliper	314	240	<0.01
Economics			
2.5% Caliper	102	60	<0.01
5% Caliper	174	140	<0.05
10% Caliper	166	130	
15% Caliper	431	429	
20% Caliper	654	605	
Non-Economics			
2.5% Caliper	79	63	
5% Caliper	120	122	
10% Caliper	234	250	
15% Caliper	348	366	
20% Caliper	481	526	

This table shows the number of results of studies that fall into each caliper, by category (RCT or non-RCT; economics or non-economics). All else equal we might expect fewer results over the caliper than under the caliper, especially for wide calipers, given the overall distribution of results.

Table 4: Caliper Tests: By Result Over Time, RCTs vs. Non-RCTs

RCTs				Non-RCTs			
	Over	Under	p-value		Over	Under	p-value
1990-1999							
2.5% Caliper	18	18		2.5% Caliper	6	3	
5% Caliper	26	29		5% Caliper	13	8	
10% Caliper	51	60		10% Caliper	23	18	
15% Caliper	77	79		15% Caliper	41	30	
20% Caliper	98	104		20% Caliper	55	47	
2000-2009							
2.5% Caliper	81	54	<0.05	2.5% Caliper	34	11	<0.001
5% Caliper	116	112		5% Caliper	50	30	<0.05
10% Caliper	222	239		10% Caliper	74	64	
15% Caliper	324	374	<0.10	15% Caliper	109	103	
20% Caliper	447	510	<0.05	20% Caliper	159	155	
2010+							
2.5% Caliper	34	34		2.5% Caliper	6	3	
5% Caliper	61	65		5% Caliper	17	8	
10% Caliper	125	132		10% Caliper	25	12	<0.05
15% Caliper	185	183		15% Caliper	37	20	<0.05
20% Caliper	269	270		20% Caliper	98	36	<0.001

This table shows the number of results of experimental and quasi-experimental studies that fall into each caliper over time. 165 results in the data with full information about levels of significance were from pre-1990 papers and not included above; 18 of these fall within the largest caliper.

5 Conclusion

This paper finds that studies using randomized experiments exhibit less specification searching and publication bias than those that do not. Papers written by researchers in economics-related disciplines also exhibit higher levels of specification searching and publication bias, though much of this appears related to the methods used. However, these biases are less pronounced than has previously been found in some of the other social sciences. The data include results from both published and unpublished papers, and unpublished papers show these discrete jumps slightly less than results from published papers.

A second contribution is that specification searching and publication bias are shown to

Table 5: Caliper Tests: By Result Over Time, Economics vs. Non-Economics

Economics				Non-Economics			
	Over	Under	p-value		Over	Under	p-value
1990-1999							
2.5% Caliper	5	3		2.5% Caliper	19	18	
5% Caliper	12	7		5% Caliper	27	30	
10% Caliper	22	16		10% Caliper	52	62	
15% Caliper	39	28		15% Caliper	79	81	
20% Caliper	51	44		20% Caliper	102	107	
2000-2009							
2.5% Caliper	67	28	<0.001	2.5% Caliper	48	37	
5% Caliper	95	69	<0.10	5% Caliper	71	73	
10% Caliper	157	158		10% Caliper	139	145	
15% Caliper	220	246		15% Caliper	213	231	
20% Caliper	303	331		20% Caliper	303	334	
2010+							
2.5% Caliper	30	29		2.5% Caliper	10	8	
5% Caliper	58	54		5% Caliper	20	19	
10% Caliper	110	105		10% Caliper	40	39	
15% Caliper	170	153		15% Caliper	52	50	
20% Caliper	298	228	<0.01	20% Caliper	69	78	

This table shows the number of results of studies by authors in economics and non-economics departments that fall into each caliper over time. 165 results in the data with full information about levels of significance were from pre-1990 papers and not included above; 18 of these fall within the largest caliper.

not be static, but biases that evolve. In particular, quasi-experimental studies have 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 (2017). "AidGrade Impact Evaluation Data, Version 1.4".

AidGrade (2017). "AidGrade Process Description".

Ashenfelter, Orley, Colm Harmon and Hessel Oosterbeek (1999). "A Review of Estimates of the Schooling/Earnings Relationship, with Tests for Publication Bias", Labour Economics, vol. 6 (4).

Bastardi, Anthony, Eric Luis Uhlmann and Lee Ross (2011). "Wishful Thinking: Belief, Desire, and the Motivated Evaluation of Scientific Evidence", Psychological Science.

Begg, Colin and Jesse Berlin (1988). "Publication Bias: A Problem in Interpreting Medical Data", Journal of the Royal Statistical Society. Series A.

Berlin, Jesse, Colin Begg and Thomas Louis (1987). "An Assessment of Publication Bias Using a Sample of Published Clinical Trials", J. Am. Stat. Assoc., vol. 84.

Brodeur, Abel *et al.* (2012). "Star Wars: The Empirics Strike Back", working paper.

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

DellaVigna, Stefano and Devin Pope (2016). "Predicting Experimental Results: Who Knows What?", working paper.

Dickersin, Kay (1990). "The Existence of Publication Bias and Risk Factors for Its Occurrence", JAMA, vol. 263.

Easterbrook, PJ *et al.* (1991). "Publication Bias in Clinical Research", Lancet, vol. 337.

Ferguson, Christopher and Michael Brannick (2012). "Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses." Psychological Methods, vol. 17 (1), Mar 2012, 120-128.

Franco, Annie, Neil Malhotra and Gabor Simonovits (2014). "Publication Bias in the Social Sciences: Unlocking the File Drawer", Science, vol. 345 (6203), Sep 2014, 1502-1505.

Gerber, Alan and Neil Malhotra (2008a). "Do Statistical Reporting Standards Affect

What Is Published? Publication Bias in Two Leading Political Science Journals", Quarterly Journal of Political Science, vol 3.

Gerber, Alan and Neil Malhotra (2008b). "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results", Sociological Methods &Research, vol. 37 (3).

Greenwald, Anthony (1975). "Consequences of Prejudice Against the Null Hypothesis", Psychological Bulletin.

Groh, Matthew, Krishnan, Nandini, McKenzie, David and Tara Vishwanath (2012). "Soft Skills or Hard Cash? The Impact of Training and Wage Subsidy on Female Youth Employment in Jordan", Policy Research Working Paper No. 6141.

Page, Matthew, McKenzie, Joanne and Andrew Forbes (2013). "Many Scenarios Exist for Selective Inclusion and Reporting of Results in Randomized Trials and Systematic Reviews", Journal of Clinical Epidemiology, vol. 66 (5).

Rosenthal, Robert (1979). "The File Drawer Problem and Tolerance for Null Results", Psychological Bulletin, vol. 86 (3).

Simes, John (1986). "Publication Bias: The Case for an International Registry of Clinical Trials", American Society of Clinical Oncology.

Simmons, Joseph and Uri Simonsohn (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant", Psychological Science, vol. 22.

Vivalt, Eva (2016). "How Much Can We Generalize From Impact Evaluations?", Working Paper.

Appendices

Guide to Appendices

Appendices in this Paper

A) Additional results.

B) Excerpt from AidGrade's Process Description (2017) and description of additional paper screening.

C) Description of data extraction for IFLS and ICRISAT papers.

Further Online Appendices

Having to describe data from twenty different meta-analyses and systematic reviews, we must rely in part on online appendices. The following are available at http://www.evavivalt.com/appendices-bias:

D) The search terms and inclusion criteria for each topic.

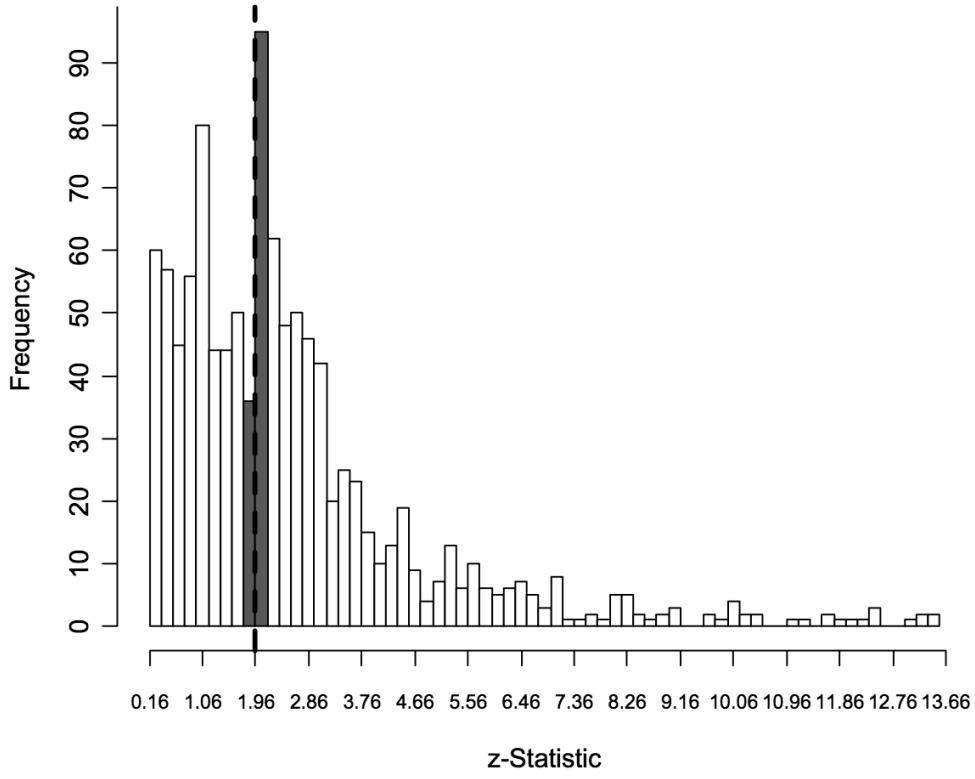
E) Bibliography of included and excluded papers.

F) The coding manual.

Appendix A

Additional Results

Figure A.1: The Distribution of z-statistics in Political Science, Reproduced from Gerber and Malhotra (2008a)



This figure plots the distribution of z-statistics found in Gerber and Malhotra (2008a). Each bar represents a 0.2 range of z-statistics, which they consider an approximate 10% caliper. The dashed line in the figure is drawn at z=1.96.

Table A.1: Caliper Tests for Political Science, Reproduced from Gerber and Malhotra (2008a)

	Over Caliper	Under Caliper	p-value
A. APSR			
Vol. 89-101			
10% Caliper	49	15	<0.001
15% Caliper	67	23	<0.001
20% Caliper	83	33	<0.001
Vol. 96-101			
10% Caliper	36	11	<0.001
15% Caliper	46	17	<0.001
20% Caliper	55	21	<0.001
Vol. 89-95			
10% Caliper	13	4	0.02
15% Caliper	28	12	0.008
20% Caliper	21	6	0.003
B. AJPS			
Vol. 39-51			
10% Caliper	90	38	<0.001
15% Caliper	128	66	<0.001
20% Caliper	165	95	<0.001
Vol. 46-51			
10% Caliper	56	25	<0.001
15% Caliper	80	45	0.001
20% Caliper	105	66	0.002
Vol. 39-45			
10% Caliper	34	13	0.002
15% Caliper	48	21	<0.001
20% Caliper	60	29	<0.001

This table is reproduced from Gerber and Malhotra (2008a) for the sake of comparison.

Table A.2: Caliper Tests: By Result, Method and Discipline

RCT				Non-RCT			
	Over	Under	p-value		Over	Under	p-value
Economic				Economic			
2.5% Caliper	60	44		2.5% Caliper	42	16	<0.001
5% Caliper	92	89		5% Caliper	74	41	<0.01
10% Caliper	176	200		10% Caliper	115	80	<0.05
15% Caliper	259	293		15% Caliper	172	136	<0.05
20% Caliper	369	401		20% Caliper	285	204	<0.001
Non-Economic				Non-Economic			
2.5% Caliper	75	62		2.5% Caliper	4	1	
5% Caliper	113	117		5% Caliper	7	5	
10% Caliper	225	235		10% Caliper	9	15	
15% Caliper	331	347		15% Caliper	17	19	
20% Caliper	452	490		20% Caliper	29	36	

This table demonstrates that there are very few non-randomized studies in this data set that were done outside of economics and divides the economics results by the method each paper used; non-randomized studies exhibited much larger biases than randomized studies.

Table A.3: Caliper Tests: By Paper

	Over Caliper	Under Caliper	p-value
All studies			
2.5% Caliper	83	59	<0.10
5% Caliper	111	97	
10% Caliper	131	141	
15% Caliper	157	166	
20% Caliper	160	192	<0.10
RCTs			
2.5% Caliper	59	50	
5% Caliper	81	79	
10% Caliper	98	114	
15% Caliper	114	132	
20% Caliper	112	151	<0.05
Non-RCTs			
2.5% Caliper	24	9	<0.05
5% Caliper	30	18	
10% Caliper	33	27	
15% Caliper	43	34	
20% Caliper	48	41	
Economic			
2.5% Caliper	34	27	
5% Caliper	53	39	
10% Caliper	58	58	
15% Caliper	69	72	
20% Caliper	67	78	
Non-Economic			
2.5% Caliper	49	32	<0.10
5% Caliper	58	58	
10% Caliper	73	83	
15% Caliper	88	94	
20% Caliper	93	109	

This figure shows the number of results of studies that fall into each caliper, by category (RCT or non-RCT; economics or non-economics). These results are aggregated to the paper level. All else equal we might expect fewer results over the caliper than under the caliper, especially for wide calipers, given the overall distribution of results.

Table A.4: Caliper Tests: By Result, Unpublished vs. Published

Unpublished				Published			
	Over	Under	p-value		Over	Under	p-value
All studies				All studies			
2.5% Caliper	26	27		2.5% Caliper	155	96	<0.001
5% Caliper	53	44		5% Caliper	233	208	
10% Caliper	95	100		10% Caliper	430	430	
15% Caliper	143	150		15% Caliper	636	645	
20% Caliper	228	212		20% Caliper	907	919	
RCTs				RCTs			
2.5% Caliper	11	21		2.5% Caliper	124	85	<0.01
5% Caliper	22	30		5% Caliper	183	176	
10% Caliper	51	67		10% Caliper	350	368	
15% Caliper	81	94		15% Caliper	509	546	
20% Caliper	145	133		20% Caliper	676	758	<0.05
Non-RCTs				Non-RCTs			
2.5% Caliper	15	6	<0.10	2.5% Caliper	31	11	<0.01
5% Caliper	31	14	<0.05	5% Caliper	50	32	<0.10
10% Caliper	44	33		10% Caliper	80	62	
15% Caliper	62	56		15% Caliper	127	99	<0.10
20% Caliper	83	79		20% Caliper	231	161	<0.001

This table presents the main findings disaggregated by publication status.

Appendix B

Description of Data Collection Process

Data from a non-profit research institute, AidGrade, were used for this paper. The following details of AidGrade's data collection process are excerpted from AidGrade's Process Description, which governed the collection of these data.

Excerpt from AidGrade's Process Description

Description of AidGrade's Methodology

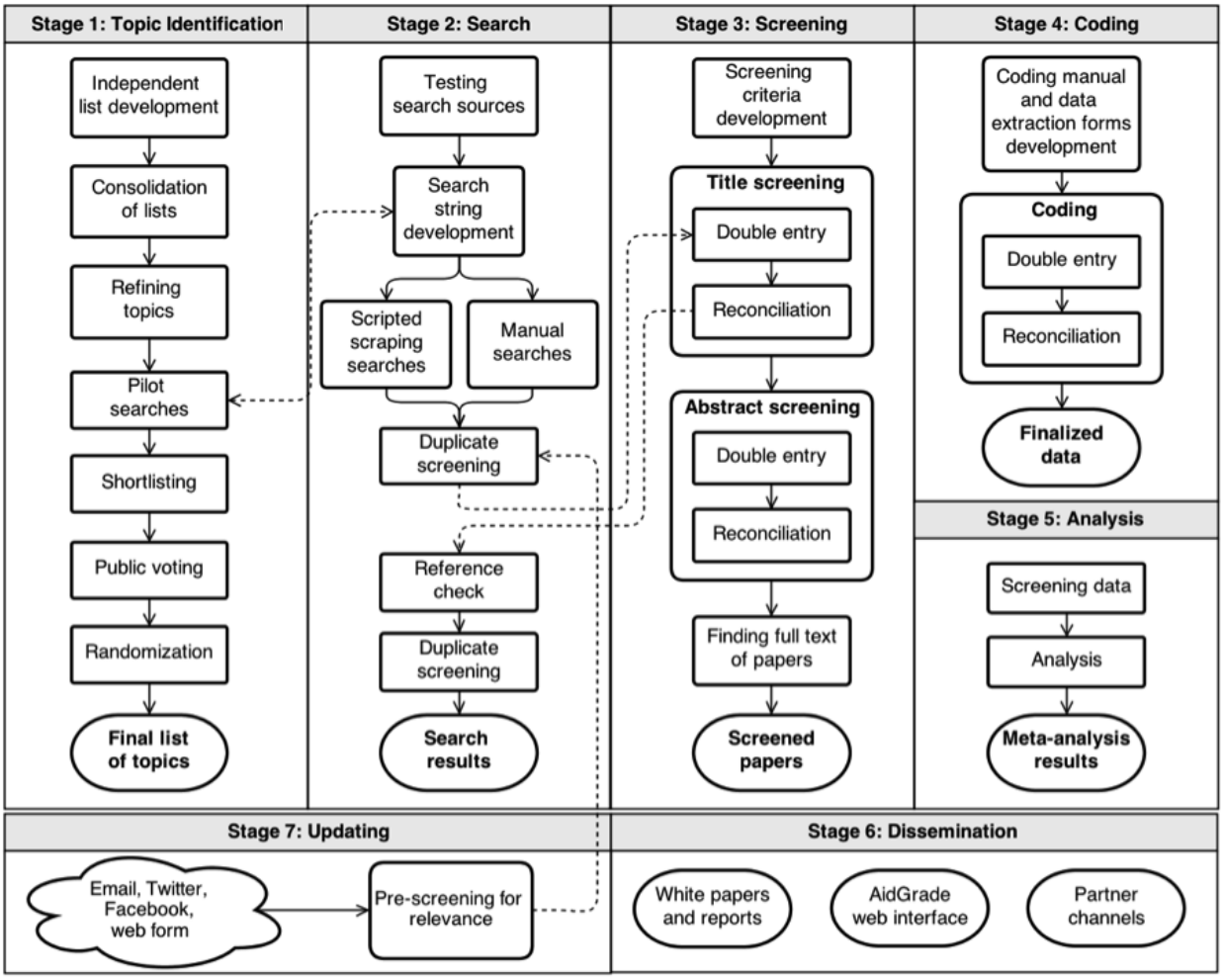
Stage 1: Topic Identification

AidGrade staff members were asked to each independently make a list of at least thirty international development programs that they considered to be the most interesting. The independent lists were appended into one document and duplicates were tagged and removed. Each of the remaining topics was discussed and refined to bring them all to a clear and narrow level of focus. Pilot searches were conducted to get a sense of how many impact evaluations there might be on each topic, and all the interventions for which the very basic pilot searches identified at least two impact evaluations were shortlisted. A random subset of the topics was selected, also acceding to a public vote for the most popular topic.

Stage 2: Search

Each search engine has its own peculiarities. In order to ensure all relevant papers and few irrelevant papers were included, a set of simple searches was conducted on different potential search engines. First, initial searches were run on AgEcon; British Library for Development Studies (BLDS); EBSCO; Econlit; Econpapers; Google Scholar; IDEAS; JOLISPlus; JSTOR; Oxford Scholarship Online; Proquest; PubMed; ScienceDirect; SciVerse; SpringerLink; Social Science Research Network (SSRN); Wiley Online Library; and the World Bank eLibrary. The list of potential search engines was compiled broadly

Figure B.1: Process Description



from those listed in other systematic reviews. The purpose of these initial searches was to obtain information about the scope and usability of the search engines to determine which ones would be effective tools in identifying impact evaluations on different topics. External reviews of different search engines were also consulted, such as a Falagas et al. (2008) study which covered the advantages and differences between the Google Scholar, Scopus, Web of Science and PubMed search engines.

Second, searches were conducted for impact evaluations of two test topics: deworming and toilets. EBSCO, IDEAS, Google Scholar, JOLISPlus, JSTOR, Proquest, PubMed, ScienceDirect, SciVerse, SpringerLink, Wiley Online Library and the World Bank eLibrary

were used for these searches. 9 search strings were tried for deworming and up to 33 strings for toilets, with modifications as needed for each search engine. For each search the number of results and the number of results out of the first 10-50 results which appeared to be impact evaluations of the topic in question were recorded. This gave a better sense of which search engines and which kinds of search strings would return both comprehensive and relevant results. A qualitative assessment of the search results was also provided for the Google Scholar and SciVerse searches.

Finally, the online databases of J-PAL, IPA, CEGA and 3ie were searched. Since these databases are already narrowly focused on impact evaluations, attention was restricted to simple keyword searches, checking whether the search engines that were integrated with each database seemed to pull up relevant results for each topic.

Ultimately, Google Scholar and the online databases of J-PAL, IPA, CEGA and 3ie, along with EBSCO/PubMed for health-related interventions, were selected for use in the full searches.

After the interventions of interest were identified, search strings were developed and tested using each search source. Each search string included methodology-specific stock keywords that narrowed the search to impact evaluation studies, except for the search strings for the J-PAL, IPA, CEGA and 3ie searches, as these databases already exclusively focus on impact evaluations.

Experimentation with keyword combinations in stages 1.4 and 2.1 was helpful in the development of the search strings. The search strings could take slightly different forms for different search engines. Search terms were tailored to the search source, and a full list is included in an appendix.

C# was used to write a script to scrape the results from search engines. The script was programmed to ensure that the Boolean logic of the search string was properly applied within the constraints of each search engines capabilities.

Some sources were specialized and could have useful papers that do not turn up in

simple searches. The papers listed on J-PAL, IPA, CEGA and 3ies websites are a good example of this. For these sites, it made more sense for the papers to be manually searched and added to the relevant spreadsheets. After the automated and manual searches were complete, duplicates were removed by matching on author and title names.

During the title screening stage, the consolidated list of citations yielded by the scraped searches was checked for any existing meta-analyses or systematic reviews. Any papers that these papers included were added to the list. With these references added, duplicates were again flagged and removed.

Stage 3: Screening

Generic and topic-specific screening criteria were developed. The generic screening criteria are detailed below, as is an example of a set of topic-specific screening criteria.

The screening criteria were very inclusive overall. This is because AidGrade purposely follows a different approach to most meta-analyses in the hopes that the data collected can be re-used by researchers who want to focus on a different subset of papers. Their motivation is that vast resources are typically devoted to a meta-analysis, but if another team of researchers thinks a different set of papers should be used, they will have scour the literature and recreate the data from scratch. If the two groups disagree, all the public sees are their two sets of findings and their reasoning for selecting different papers. AidGrade instead strives to cover the superset of all impact evaluations one might wish to include along with a list of their characteristics (*e.g.* where they were conducted, whether they were randomized by individual or by cluster, *etc.*) and let people set their own filters on the papers or select individual papers and view the entire space of possible results.

Figure B.2: Generic Screening Criteria

Category	**Inclusion Criteria**	**Exclusion Criteria**
Methodologies	Impact evaluations that have counterfactuals	Observational studies, strictly qualitative studies
Publication status	Peer-reviewed or working paper	N/A
Time period of study	Any	N/A
LocationGeography	Any	N/A
Quality	Any	N/A

Figure B.3: Topic-Specific Criteria Example: Formal Banking

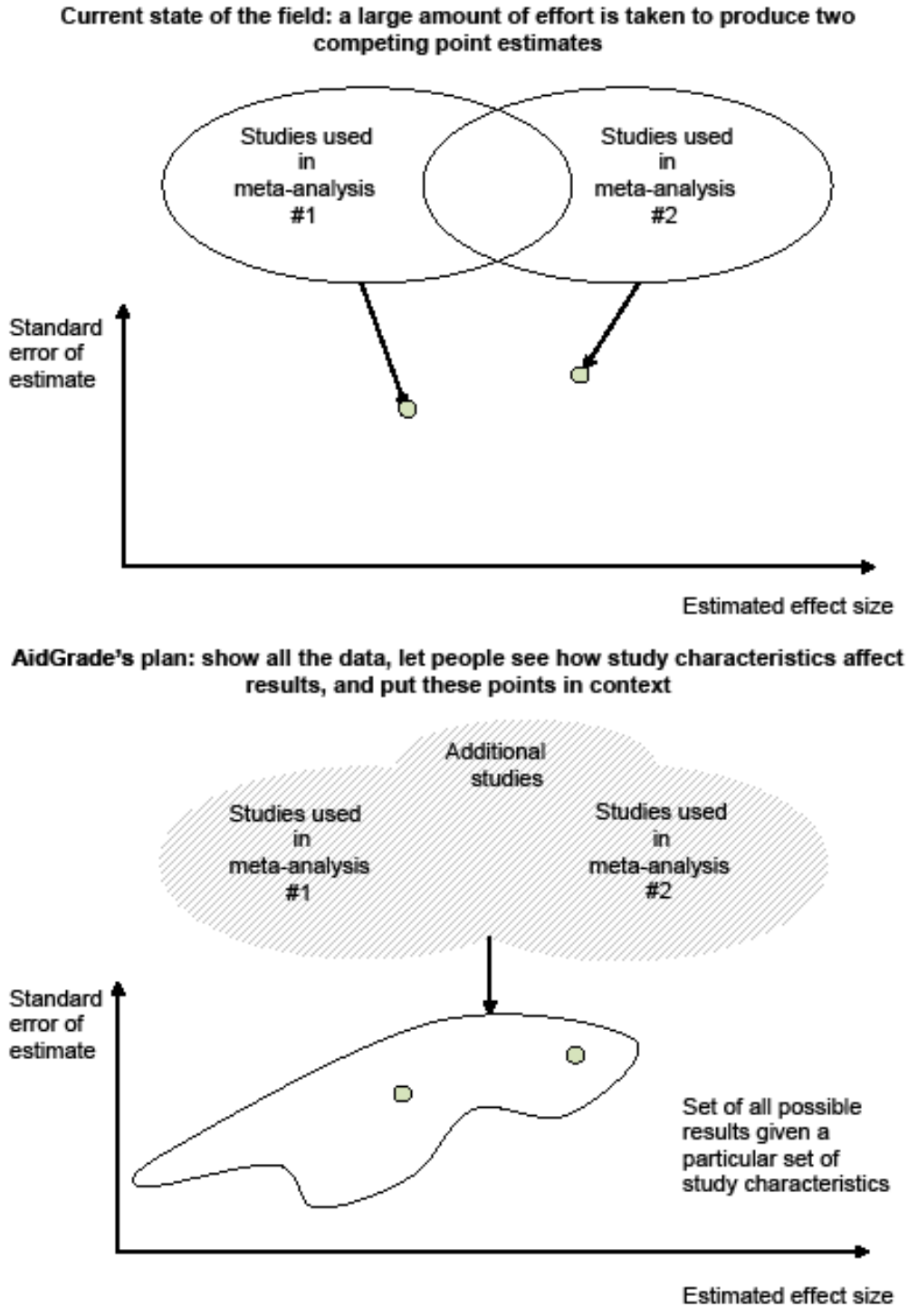
Category	**Inclusion Criteria**	**Exclusion Criteria**
Intervention	Formal banking services specifically including: - Expansion of credit and/or savings - Provision of technological innovations - Introduction or expansion of financial education, or other program to increase financial literacy or awareness	Other formal banking services Microfinance
Outcomes	- Individual and household income - Small and micro-business income - Household and business assets - Household consumption - Small and micro-business investment - Small, micro-business or agricultural output - Measures of poverty - Measures of well-being or stress - Business ownership - Any other outcome covered by multiple papers	N/A

Figure 12 illustrates the difference.

For this reason, minimal screening was done during the screening stage. Instead, data was collected broadly and re-screening was allowed at the point of doing the analysis. This is highly beneficial for the purpose of this paper, as it allows us to look at the largest possible set of papers and all subsets.

After screening criteria were developed, two volunteers independently screened the titles to determine which papers in the spreadsheet were likely to meet the screening criteria developed in Stage 3.1. Any differences in coding were arbitrated by a third volunteer. All

Figure B.4: AidGrade's Strategy



volunteers received training before beginning, based on the AidGrade Training Manual and a test set of entries. Volunteers' training inputs were screened to ensure that only proficient volunteers would be allowed to continue. Of those papers that passed the title screening, two volunteers independently determined whether the papers in the spreadsheet met the screening criteria developed in Stage 3.1 judging by the paper abstracts. Any differences in coding were again arbitrated by a third volunteer. The full text was then found for those papers which passed both the title and abstract checks. Any paper that proved not to be a relevant impact evaluation using the aforementioned criteria was discarded at this stage.

Stage 4: Coding

Two AidGrade members each independently used the data extraction form developed in Stage 4.1 to extract data from the papers that passed the screening in Stage 3. Any disputes were arbitrated by a third AidGrade member. These AidGrade members received much more training than those who screened the papers, reflecting the increased difficulty of their work, and also did a test set of entries before being allowed to proceed. The data extraction form was organized into three sections: (1) general identifying information; (2) paper and study characteristics; and (3) results. Each section contained qualitative and quantitative variables that captured the characteristics and results of the study.

Stage 5: Analysis

A researcher was assigned to each meta-analysis topic who could specialize in determining which of the interventions and results were similar enough to be combined. If in doubt, researchers could consult the original papers. In general, researchers were encouraged to focus on all the outcome variables for which multiple papers had results.

When a study had multiple treatment arms sharing the same control, researchers would check whether enough data was provided in the original paper to allow estimates to be combined before the meta-analysis was run. This is a best practice to avoid double-counting

the control group; for details, see the Cochrane Handbook for Systematic Reviews of Interventions (2011). If a paper did not provide sufficient data for this, the researcher would make the decision as to which treatment arm to focus on. Data were then standardized within each topic to be more comparable before analysis. Units were converted where possible. The standardized mean difference was also calculated using the pooled standard deviation of the outcome variable across the treatment and control groups where available. Where this measure was not available, the standard deviation in the control group was preferentially used, followed by the standard deviation in the treatment group, followed by the standard deviation of the outcome variable from other studies.

The standardized mean difference can also be approximated by $\sqrt{(3)}/\pi * \ln$ odds ratio (Higgins and Green, 2011). We conducted these data transformations where it would preserve more data. The limitations in what papers reported meant that sometimes intervention-outcome combinations that had sufficient unstandardized data for analysis did not have sufficient standardized data for analysis, and vice versa.

The subsequent steps of the meta-analysis process are irrelevant for the purposes of this paper. It should be noted that the first set of ten topics followed a slightly different procedure for stages (1) and (2). Only one list of potential topics was created in Stage 1.1, so Stage 1.2 (Consolidation of Lists) was only vacuously followed. There was also no randomization after public voting (Stage 1.7) and no scripted scraping searches (Stage 2.3), as all searches were manually conducted using specific strings. A different search engine was also used: SciVerse Hub, an aggreator that includes SciVerse Scopus, MEDLINE, PubMed Central, ArXiv.org, and many other databases of articles, books and presentations. The search strings for both rounds of meta-analysis, manual and scripted, are detailed in an online appendix.

Stage 6: Updating

Data are subject to periodic updating. Unlike a static database, AidGrade's database is

intended as a living database. Research assistants add papers to the database as they are brought to AidGrade's attention, such as by authors e-mailing AidGrade their papers. The same screening criteria and data extraction forms are used.

To ensure replicability of results, AidGrade's database is versioned.

Appendix C

Description of Data Extraction for IFLS and ICRISAT Papers

In addition to AidGrade's database of impact evaluation results, this study considered results from papers using the IFLS or ICRISAT daa sets.

Lists of papers using IFLS or ICRISAT data are independently maintained by outside parties (ICRISAT: http://vdsa.icrisat.ac.in/vdsa-jarticles.htm; IFLS: http://www.rand.org/labor/FLS/IFLS/papers.html). We used these lists of papers rather than conducting our own searches, using the papers that were listed as of February, 2014 for the ICRISAT papers and April, 2014 for the IFLS papers.

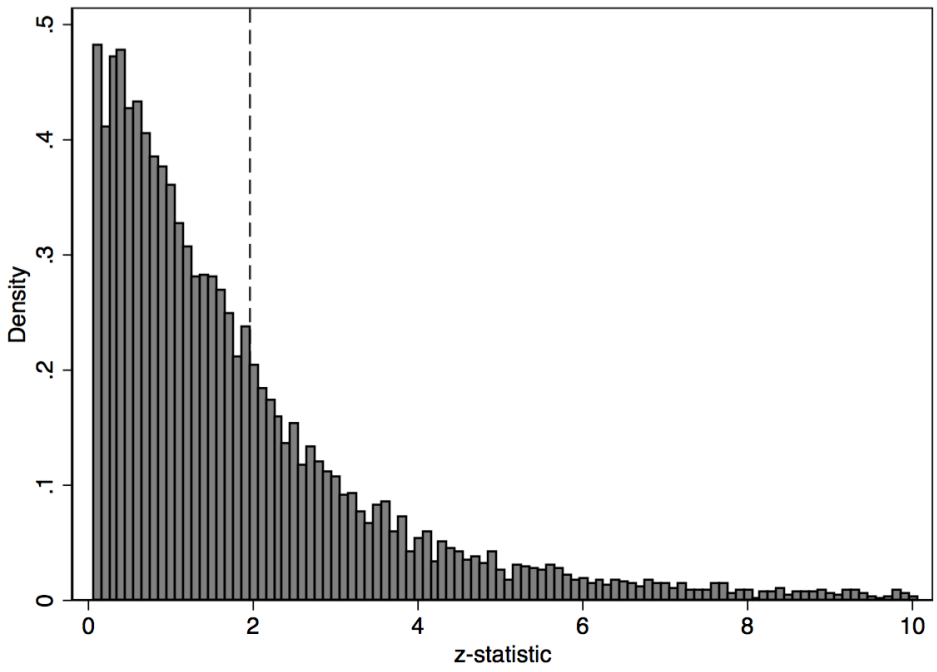
Some papers that were cited at those links were unable to be located. In total, we collected data for 116 supplementary papers, using the same data extraction templates as used by AidGrade for its meta-analyses and systematic reviews. However, on close examination, it was determined that most of the data came from regressions reporting correlations. Not only was there rarely an experimental or quasi-experimental design, but the purpose of the regressions often did not seem to be to test a particular hypothesis; it should be noted that when many of these papers were written, it was common to run "kitchen sink" style regressions containing many variables. The authors may later point to the regression results that turned out to be significant, but the point often seemed not to be to test a particular hypothesis but to present exploratory analyses.

These data are interesting, but may not be the most comparable to the studies in AidGrade's database. Thus, the great majority of results were later excluded from the analyses in this paper. In total, results from only 33 papers were used: those which were determined by coders to be explicitly testing a hypothesis, according to the paper's text.

The distribution of the z-statistics across the other 10,747 results that were not considered in this paper is presented below. As these results were essentially exploratory and threw in a "kitchen sink" of potentially explanatory variables, there is no obvious

pattern of specification searching or publication bias. That these results were not focused on explicitly testing hypotheses is also evident from their sheer number: the median paper presented 57 such "results", not counting results from alternative specifications, compared to a median of 10 results testing specific hypotheses in the main data set.

Figure C.1: Distribution of z-statistics in Exploratory Analyses



This figure shows the distribution of z-statistics for exploratory analyses in the IFLS and ICRISAT papers. These analyses were not, according to each paper's text, specifically attempting to test a particular hypothesis. While one might expect authors to "fish" for results even in this context, it appears that even if authors may have appreciated significant results these significant results were outweighed by the vast number of insignificant correlations.