Essays on generalizability and evidence-use in policy

Michelle Rao

A thesis presented for the degree of Doctor of Philosophy.

Department of Economics

London School of Economics

London, United Kingdom

August 28, 2025

Declaration

Statement of inclusion of previous work

I confirm that Chapter 2 of this thesis is a revised version of the paper I submitted at the end of my Masters of Research in 2020. Chapter 3 is joint work with Oriana Bandiera, Barbara Petrongolo, and Nidhi Parekh. It was published on January 2022, in *Economica*, Volume 89, Issue S1.

Statement of conjoint work

Chapter 3 is joint work with Oriana Bandiera, Barbara Petrongolo, and Nidhi Parekh. We all contributed equally to this work.

Abstract

To increase evidence-use in policy, it is important to understand both the generalizability of evidence, and the existing use of evidence for policy decisions. This thesis comprises three papers on this topic. Chapter 1 studies evidence-use in policy, focusing on one of the most heavily evaluated anti-poverty programs — Conditional Cash Transfers (CCTs). Using a novel dataset of 128 program evaluations of CCTs in Latin America and the Caribbean mapped to policy spending on the evaluated programs, I find a robust zero relationship between research results and spending. The only exception is when evaluations are timely and politically aligned. When evaluations are released within four years of the effect year and can be attributed to the political party in power, there is a positive and significant relationship between evaluation outcomes and spending. Chapters 2 and 3 examine the generalizability of evidence, using Bayesian hierarchical models to aggregate the evidence-base on gender differences in altruism and overconfidence. In chapter 2, I find that women give three percentage points more than men in dictator games, but this estimate is likely to be an upper bound due to publication bias. In chapter 3, joint with Oriana Bandiera, Barbara Petrongolo, and Nidhi Parkeh, we find that while experts believe that men are overconfident and women are underconfident, the literature suggests that both men and women are overconfident.

Acknowledgements

I am immensely grateful to everyone who has supported me through my PhD journey. To my advisors, Rafe Meager, Oriana Bandiera, Gharad Bryan, and Robin Burgess: thank you for your invaluable guidance and support. This thesis would not have been possible without you.

I benefited greatly from feedback and conversations with numerous people who engaged with my work and ideas. This includes: Tim Besley, Mike Callen, Gabriela Diaz Pardo, Gaby Deschamps, Matthias Doepke, Nilmini Herath, Amen Jalal, Gilat Levy, Santiago Levy, Gabriel Leite Mariante, Canishk Naik, Steve Pischke, Ronny Razin, Jack Thiemel, Sarah Winton, Alwyn Young, and participants of the LSE development economics work in progress, SPEECH, Doctorissimes, and OxDev. For various discussions, their humour, and moral support, I am grateful to Alix Bonargent, Arnaud Dyèvre, Andrés Fajardo-Ramirez, Veronica Salazar-Restrepo, Javad Shamsi, Cecilia Wood, and everyone in the 3.12 office.

To my co-authors: thank you for working with me to study questions on generalizability and evidence-use in policy. Our work together has been central to shaping my thinking on this topic, and have made me the researcher I am today.

To Mayanka, Ashley, Peter, Andrew, Isheeta, Aaron, and Mickaël: thank you for keeping me grounded and (relatively) sane throughout this process, whether that was through impromptu phone calls, shared dinners, climbing sessions, board games, or pints at the Dove. To Linda and Chris: thank you for always putting things into perspective, with a perfectly timed visit or an expertly crafted e-mail or text. To Corrie: thank you for your guidance and for fostering such a warm and thoughtful community of ashtangis.

To Steph, Jacob, and my parents: thank you for your unconditional love and support. Mom and Dad, thank you for instilling in me a love for learning, and for always celebrating me for who I am. Steph, thank you for being my fiercest cheerleader, and the best sister anyone could ever ask for. Jacob, thank you for being my rock.

Contents

List of Tables

List of Figures

Chapter I

Program Evaluations and Policy Spending

Abstract: Program evaluations are motivated in part by a desire to improve policy effectiveness. Yet there is limited empirical evidence on the efficacy of evaluation itself. This paper examines the systematic relationship between program evaluations and changes in policy spending, in the context of Conditional Cash Transfers in Latin America and the Caribbean. Using a novel dataset of 128 program evaluations mapped to spending on the corresponding evaluated programs, I find a robust zero relationship between research results and spending. This holds for several definitions of evaluation outcomes: more statistically significant, larger magnitude, more surprising, or more positively framed results, do not correspond with larger increases in spending. As policymakers may learn from cumulative evidence rather than individual studies, I then use a Bayesian hierarchical approach to aggregate evaluations. I find a zero association between a country's cumulative evidence base and its spending. Finally, I explore mechanisms for this result by considering heterogeneous responses to evaluations that are more credible, actionable, or generalizable. I find that credibility and generalizability are unrelated to spending, but evaluations conducted quickly (within four years of the effect year) and attributable to the political party in power, are significantly predictive of spending. Thus, timeliness may be an overlooked aspect of the evidence-to-policy pipeline.

I.1 Introduction

Program evaluations are increasingly common in policy settings, with governments and international institutions playing an active role in advocating for, funding, and conducting evaluations (Levine and Savedoff, 2015; Independent Evaluation Group, 2012; USAID, 2016). However, there is limited evidence on the empirical relationship between the

results of these evaluations and key policy decisions. In providing causal estimates of impact, program evaluations can - in theory - have a direct impact on policy decisions such as policy spending, program design, and program adoption (Duflo and Banerjee, 2011). Yet, the applicability of evidence to policy decisions can also be limited by features of the political environment, or of the evidence-base itself (e.g. Allcott, 2015; Rosenzweig and Udry, 2020). Understanding the relationship between evidence and policy is a fundamental step to maximising the policy impact of research.

This paper contributes to this understanding by exploring the relationship between program evaluations and policy spending, in the context of Conditional Cash Transfers (CCTs) in Latin America and the Caribbean. The early studies of Mexico's PROGRESA (e.g. Gertler, 2004; Schultz, 2004) spurred the rise of a culture of evaluation of CCTs, particularly for countries in the region (Fiszbein and Schady, 2009). Between 2000 to 2015, there were 31 evaluated CCTs across 17 countries in Latin America and the Caribbean. CCTs are therefore often cited as a success story for evidence-based policy, with one narrative being that evaluation results influenced spending decisions by helping to direct resources into programs with higher proven impact (e.g. Angrist and Pischke, 2010; Duflo and Kremer, 2003). However, little is known about the empirical relationship between evaluation outcomes and policy spending decisions.

This is the focus of my study. I examine the relationship between program evaluations and changes in policy spending for CCTs in Latin America and the Caribbean, from 2000 to 2015. I construct a novel dataset of program evaluations of CCTs mapped to annual spending on the evaluated programs. My dataset covers a total of 128 program evaluations[1], representing 468 headline results on the causal impact of CCTs on poverty-related outcomes. Using this data, I examine patterns of evidence-based policy spending consistent with two broad categories of evidence-use: immediate spending responses to findings from individual evaluations; and gradual spending responses to the aggregate evidence-base. Lastly, I examine features of evidence that may matter for pol-

[1]I define program evaluations as studies that estimate the causal impact of being a program recipient, compared to a relevant counterfactual of being a non-recipient. This includes studies that use experimental and/or non-experimental methods to estimate the causal effect of interest.

icy decisions, using variation in evaluation characteristics associated with higher policy relevance.

To study this relationship empirically, I first define what patterns in the data would be consistent with evidence-based policy spending. Using a simple theoretical model of policy-making under uncertainty, I show that, under basic assumptions on evidence quality, evidence-based policy spending would be observed empirically as a positive relationship between evaluation outcomes and spending if: (1) policymakers use evidence to update their beliefs; and (2) the perceived benefits of adjusting policy spending outweigh the costs to policy change. The relationship between research findings and spending therefore depends not only on the evidence-base, but also its interaction with political and other costs.

There are two challenges to discerning this relationship in the data that I address in my empirical strategy. First, even if policymakers are using evidence, I cannot observe the subset of evaluations – the information set – that policymakers use to make spending decisions. Even within a country, CCT programs are repeatedly evaluated.[2] As a result, policymakers could be learning from either individual evaluations, or from the cumulative set of evaluations on their program. Second, even given a fixed information set, I cannot observe what findings policymakers extract, and how they process the results of the evaluation. Thus, studying the relationship between evaluation outcomes and spending requires careful aggregation of evaluation findings both within and across studies. I therefore consider the relationship between program evaluations and spending for aggregations of evaluation findings for individual evaluations, and for cumulative evidence from each country.

In the first part of the paper, I consider the relationship between individual evaluation outcomes and policy spending on the evaluated CCTs. Using reported treatment effects from individual evaluations, I find that larger estimates of impact do not correspond with larger changes in spending on the evaluated program. The zero relationship

[2] Almost all countries in Latin America and the Caribbean have had more than three program evaluations on the impact of a CCT in their country, from 2000-2015. The median country has had seven CCT program evaluations over this time period.

holds regardless of the way in which I summarise reported treatment effects from each evaluation. There is no association between spending and the statistical significance of treatment effects, as captured by the mean or maximum of the precision-weighted treatment effect (i.e. the t-statistic) of headline results. There is also no association between spending and the magnitude of treatment effects, as captured by the mean or maximum effect size of headline results. The estimated relationship between treatment effects and changes in policy spending is statistically insignificant and economically small in magnitude. Compared with an evaluation that finds a null result, a positive and significant evaluation would be associated with a 1.65 million USD increase in spending, which accounts for less than 1% of the average annual change in spending.

One limitation of the baseline relationship between reported treatment effects and spending is that reported outcomes do not account for policymaker's prior beliefs on policy effectiveness. If policymakers have evidence-based priors, a zero association between spending and evaluation outcomes that are aligned with the existing evidence-base would be consistent with evidence-use. Using a fixed-effects model to aggregate findings, I estimate time- and country-specific prior beliefs on the effectiveness of CCTs. I find that more surprising findings – those that deviate more from these evidence-based priors – do not correspond to larger changes in spending. Evaluations that have more positive results, relative to the existing evidence base, do not correspond to larger increases in spending. Furthermore, evaluations that have more negative results, relative to the existing evidence base, do not correspond to larger decreases in spending.[3] These results hold regardless of assumptions on the weight countries place on evidence from other countries' programs when forming their prior beliefs – that is, assumptions on the perceived external validity of evaluations from other countries.

Beyond quantitative measures of treatment effects, the strength of evaluation results is also conveyed through the language used to describe findings. Using sentiment analysis on the abstract text, I estimate how positively or negatively framed research results

[3]This is in contrast to Vivalt and Coville (2023), who find that policymakers update their beliefs more in response to good news, relative to bad news.

are. Authors tend to use more positive than negative language to describe their research findings. However, I find that more positively framed evaluations do not correspond to larger increases in policy spending.

In the second part of the paper, I expand the information set to the cumulative evidence on CCTs, to explore patterns of policy spending explained by evidence accumulation over time. While I find a robust zero relationship between individual evaluations and program spending, sophisticated users of evidence may instead learn from the aggregate evidence-base. In the presence of limited external validity[4], combining evaluation outcomes from multiple studies can increase the ability to learn about the underlying treatment effect. Thus, policymakers that place greater weight on knowledge accumulation may be more inclined to respond to the aggregate evidence-base[5]. I use tools from meta-science – increasingly used in economics – to aggregate findings from the existing body of evidence (e.g. Banerjee et al., 2015; Meager, 2019).

I aggregate findings from each country's evidence-base using a Bayesian hierarchical model. The hierarchical structure disentangles between heterogeneity across studies arising from sampling variation versus genuine variation in treatment effects. This gives an estimate of the true average that adjusts for these different sources of heterogeneity. I find that stronger aggregate evidence of the effectiveness of CCTs in each country – that is, a higher posterior mean on treatment effects – does not correspond with higher spending on CCTs. This is not because studies are not informative about the underlying population treatment effect. I estimate the generalized pooling factor from the Bayesian model (Gelman and Pardoe, 2006). I find that in most countries, there is a considerable amount of pooling across studies, indicating a reasonable amount of external validity.

Taken together, these findings show that there is a robust and relatively precise zero correlation between evaluation outcomes and subsequent spending on the evaluated pro-

[4]For instance, Allcott (2015) finds evidence of site selection bias, whereby program impacts are positively correlated with local characteristics, implying that there is limited external validity of individual program evaluations.

[5]Dunning et al. (2019) find some evidence of in support of this hypothesis. In a lab-in-field experiment with policy practitioners, they find that exposure to meta-analysis increases the accuracy of out-of-sample predictions.

gram. These results suggest that either policymakers do not adjust their spending in response to evaluation outcomes; or, there is a complex process that directly offsets any changes made, resulting in a reliable zero correlation. Lab-in-field studies show that policymakers can update their beliefs in response to research to varying degrees (e.g. Nakajima, 2021; Vivalt and Coville, 2023; Hjort et al., 2021; Banuri et al., 2017; Dunning et al., 2019). In my setting, I do not measure changes in beliefs. However, given that policymakers are highly trained, and are often directly or indirectly involved in the evaluation of CCTs, the zero relationship seems unlikely to be driven by a lack of policy awareness. Rather, my results suggest that evidence from program evaluations do not overcome the practical or political costs of changing policy spending.

To better understand the costs to evidence-use, I consider patterns of spending in response to subsets of evidence that are likely to be more policy relevant. I examine differential responsiveness along three dimensions of evidence characteristics: (1) credibility – the extent to which the evaluation gives internally valid, and reliable estimates of the causal impact of a program; (2) generalizability – the extent to which the evaluation is informative and relevant to a broader population of interest; and lastly, (3) actionability – the extent to which the evaluation gives impact estimates that are timely and embedded in the policymaker's decision process.

I find no evidence of selective responsiveness to more credible or generalizable evaluations. First, there is a zero relationship between evaluation outcomes and spending for more credible studies, as proxied by randomised controlled trials, and by studies published in top academic journals. Second, I find a zero association between research findings and subsequent spending for more generalizable studies, that measure impacts for a broader population, and for studies that are more externally valid, as proxied by a higher pooling factor from the Bayesian hierarchical model.

The only characteristic that is predictive of spending decisions is the actionability of evaluations. When evaluations are timely - available within the mean of four years after the effect year - there is a positive and significant relationship between evaluation

outcomes and changes in spending.[6] This positive association is highest for timely evaluations that have the same political party in power during the effect year and the first date of publication. These findings are suggestive of the importance of both timeliness and political alignment in evidence-use. Evaluations that are made available later relative to the effect year are likely to become less informative about current policy decisions, as the policy and economic environment changes over time. Moreover, even when evaluation results remain informative, incentives for evidence-use are likely to weaken when evaluation outcomes cannot be attributed to the current political party in power.

More broadly, my findings underscore the importance of understanding the empirical relationship between research and policy. While the literature on optimal research design often assumes that policymakers use evidence for policy decisions (e.g. Kitagawa and Tetenov, 2018; Frankel and Kasy, 2022; Haushofer et al., 2022), my findings suggest that this assumption cannot be taken as given. The positive association between evaluation outcomes and spending for timely and politically aligned evaluations is suggestive of the presence of costs to evidence-use, that may be increasing over time. Understanding these costs, along with broader aspects of the research-policy relationship, represents a valuable avenue for future research.

Most closely related to this paper are DellaVigna et al. (2024) and Wang and Yang (2021), who study policy experimentation and evidence-use in government institutions. Similar to DellaVigna et al. (2024), who study the take-up of nudges following individual experiments, I find limited evidence of responsiveness to individual evaluations. Wang and Yang (2021) study policy experimentation across states in China. They find that policy experimentation is more likely to happen in states with higher economic development, and hence there is limited scope for external validity and policy learning across states. In contrast, Hjort et al. (2021) find that randomly informing policymakers on the effectiveness of a single effective policy intervention increases the probability of adop-

[6]The timeliness of evaluation is defined as the number of years between the year of the treatment effect (i.e. endline year of data collection), and the year at which the evaluation is made available. Studies are defined as timely if the lag is less than or equal to the mean of four years. Results are robust to different definitions of 'timely' evaluations.

tion.

I contribute to this literature in two ways. First, rather than focusing on the use of evidence on multiple policies within a single institutional setting, I study evidence-use for a single policy that has been evaluated repeatedly across countries and over time. The setting of Conditional Cash Transfers means that I can explore patterns of evidence-based policy spending consistent with models of both immediate learning, from individual program evaluations, and sophisticated learning, based on the entire evidence-base. Second, I provide new evidence on policy responsiveness to research along the intensive margin of program spending. While existing studies of evidence-use within organisations focus on the extensive margin of policy take-up (e.g. Wang and Yang, 2021; DellaVigna et al., 2024), fixed costs to program setup are often very high and less likely to be comparable across contexts. Hence, marginal responses on the intensive margin of spending are an important dimension for understanding potential policy learning and evidence-use.

Lastly, I provide suggestive evidence on the features of evidence that matter for policy. Existing studies of policymaker beliefs provide insights into evidence characteristics that potentially matter for evidence-use, including the internal validity of evaluations (Mehmood et al., 2021), aspects of external validity such as sample size and country of evaluation (Hjort et al., 2021; Nakajima, 2021), the complexity of research findings (Toma and Bell, 2024). Bonargent (2024) finds evidence of higher policy implementation when projects are conducted in collaboration with policymakers. My findings suggest that the actionability of research results, and in particular – the timeliness of evaluation – is an overlooked channel to increasing the use of evidence for policy.

The rest of the paper proceeds as follows. Section I.2 describes the data and context. Section I.3 lays out the conceptual framework and empirical strategy. Section I.4 and section I.5 outlines the main results on individual evaluations and spending, and cumulative findings and spending, respectively. Section I.6 discusses the results, and implications for alternative models of evidence-based policy. Section I.7 explores heterogeneity in spending responses to different features of evidence. Section I.8 concludes.

I.2 Data & Context

Conditional Cash Transfer programs are a widespread policy instrument for poverty alleviation and are heavily studied, particularly in Latin America and the Caribbean, the focus on my study. The rigorous evaluation of Mexico's PROGRESA in the early 2000s contributed to a rise in evaluation of CCTs (Fiszbein and Schady, 2009). By 2015, Conditional Cash Transfers became a widespread policy instrument, but systematic evaluation is particularly prevalent in Latin America and the Caribbean, where almost all countries in the region had an established CCT program with an associated program evaluation (see figure I.2).

I construct a novel dataset of all program evaluations of CCTs in Latin America and the Caribbean mapped to policy spending on the same programs, between 2000 to 2015. In sections I.2.1 and I.2.2, I describe the methods used to identify the key variables in this dataset. These are broadly categorized into variables related to:

1. Program evaluations, that estimate the causal impact of CCTs on poverty-related outcomes. I describe the criteria for identifying relevant studies and relevant results of interest. I also outline here the methods used to identify key characteristics of the evaluations, including the study's origins and relationship with government;

2. Program characteristics on the evaluated programs. This includes information on policy spending, the amount spent on the evaluated CCTs, and other characteristics of the evaluated CCT.

In section I.2.3, I provide some descriptive facts and context about evaluations and spending on CCTs in Latin America and the Caribbean.

I.2.1 Program evaluations

I collect data on the estimated causal impact of Conditional Cash Transfer (CCT) programs. I focus on program evaluations of large-scale national Conditional Cash Transfer

Programs in Latin America and the Caribbean, between 2000 to 2015. The evaluated programs are institutionalised national programs for poverty alleviation, central to the country's social protection strategies.

Identifying studies: I begin by identifying relevant studies on CCTs. My starting point is the Bastagli et al. (2016) literature review on program evaluations of CCTs in lower and middle-income countries. Bastagli et al. (2016) include peer-reviewed and working papers published in academic journals and key policy-relevant grey literature (e.g. IFPRI, WB working paper) between 2000 to 2015. The studies use either experimental (e.g. Randomised Controlled Trials) or non-experimental methods (e.g. Differences in Differences, Instrumental Variables, Propensity Score Matching) to identify the causal impact of receiving a cash transfer on poverty-related outcomes in the domains of education, employment, empowerment, health, monetary poverty, and savings, investment, and production.

Importantly, I focus exclusively on studies that estimate the causal impact of being a CCT recipient, compared to a relevant counterfactual of being a non-recipient. This means that I exclude program evaluations that only compare the impact of design features (e.g. Barrera-Osorio et al., 2008). I also exclude papers that are not program evaluations, but instead use CCTs to estimate structural parameters in economic models (e.g. Attanasio and Lechene, 2010, 2014). Focusing on the subset of studies in Bastagli et al. (2016) that are in my region of study, I identify a total of 72 relevant studies across 12 (out of 23) countries in Latin America and the Caribbean.

I apply the same search criteria laid out by Bastagli et al. (2016) to identify relevant studies for the remaining 11 countries in my sample[7]. Using this search criteria, I identify an additional 20 program evaluations of CCTs in the region. I apply the same search methodology in Spanish, to identify 30 additional local language papers. Lastly, I verify my sample of studies against the GiveDirectly Cash Evidence explorer (GiveDirectly, 2023). This adds 6 studies to my sample. In total, I identify 128 relevant studies for my

[7] Bastagli et al. (2016) focus on studies taking place in lower and lower-middle income countries, as determined by the World Bank classifications in 2015. As such, studies conducted in countries like Argentina and Chile are not included in their review.

analysis.[8]

Headline Results: For each of these 128 studies, I collect data on up to six headline results on the causal impact of the cash transfer program. That is, results that are mentioned as key findings by the authors, either in the abstract or in the introduction of the paper. Many of the program evaluations in my sample run multiple regression specifications on various outcomes. By focusing on headline results, my dataset captures the key takeaways of each evaluation. For each identified headline result, I collect information on the treatment effect, the sample size, and the standard error [9]. I obtain further information on the paper's estimation strategy, the baseline and endline years pertaining to the program evaluation, and details on the sub-population for whom the treatment effect is estimated, including the gender, age range, and rural-urban classification.

This gives me a total of 128 program evaluations representing 468 headline results estimating the causal impact of CCTs poverty-related outcomes. As seen in table I.1, the headline results can be broadly classified into six outcome areas: education, health and nutrition, employment, and empowerment, monetary poverty, and savings, investment, and production. Out of 128 total studies, 50 use experimental variation to identify the causal impact of CCTs. 79 use non-experimental methods, such as propensity score matching, Differences in Difference, Regression Discontinuity and Instrumental Variables[10].

Paper characteristics: I collect data on study characteristics related to the timing and source of the program evaluation. Firstly, I identify the **earliest publication date** of the program evaluation, defined as the earliest date at which a full draft of the paper was made publicly available. Publication timelines in Economics average 16 months after submission (Hadavand et al., 2021) and researchers often share preliminary results prior to formal publication. Thus, identifying the earliest date of publication gives me a mea-

[8] See Appendix A.1 for a further breakdown of the search strategy

[9] For 36 papers in my sample, authors do not report the standard errors. In these cases, I collect relevant information needed to calculate the standard error of the main treatment effect, such as the standard deviation, the test statistic, or the p-value. If none of this information is provided, I use information on the significance of the estimate (e.g. 5% significant) to impute the largest standard error that would be correspond to the significance category.

[10] One study uses a combination of experimental and non-experimental methods in their analysis

sure of the earliest date at which research results were likely made available to policy makers.

I identify earlier versions of the papers in four steps: (1) using a citation search on google scholar, to look for earlier or later versions of the paper; (2) searching for alternative publications in IDEAS RePec; (3) keyword search of *author name + paper key words + working paper*. This helps to identify earlier or later versions of the same paper that may have a different name; and (4) search institutional or author webpages for earlier versions of the paper. For papers that are submitted in journals but that do not have an earlier version identified in the method above, I use the journal submission date as the earliest date of publication. I identify alternative publication dates for 71 of the papers in my sample.

Lastly, I collect information on the **study author** and the **origins** of paper, particularly in relation to the government. Information on both of these characteristics is often made available in the acknowledgements or notes section of the paper[11]. Using this information, I identify whether or not any of the study authors collaborated with the government at some point during the program evaluation.[12] A study is classified as having an author and institutional collaboration if the study author collaborates with the government or institution to conduct the study. I find that 65 out of 128 studies in my sample have at least one author affiliated with the governing institution.

On the **origins of the program evaluation**, I identify the demanding and evaluating agent of the program evaluation, and the relationship between the two agents. I classify demanding and evaluating agents into one of the following categories: implementing government, international institution, research centres or consultancy, or independent researcher. A study is classified as an institutional evaluation if it is demanded by either the government or international institution. A study is classified as an independent evaluation if it is both demanded and evaluated by an independent researcher.

[11] See Appendix A.1 for more detail on data collection of study characteristics

[12] If there was no information on government relationships in the paper, I search for author and government relationships related to the CCT programs using the author's public online profiles.

I.2.2 Program characteristics

I map the program evaluations of Conditional Cash Transfers to data on annual program expenditure for the same programmes. I use data from the Non-contributory Social Protection programmes in Latin America and the Caribbean database, developed by the Social Development Division of the Economic Commission for Latin America and the Caribbean (ECLAC). The database uses official country documents to report on key design characteristics of national CCT programs and, importantly for my purposes, annual budgets and expenditure on CCTs.

To capture the annual spending on conditional cash transfers, I use data reported on expenditure and budget allocations. Cecchini and Atuesta (2017) details the methodology used to harmonise the data. I use the annual budget allocations as a measure of annual spending on the CCT program, since this is the most consistently reported across the countries and over the time period of analysis. When the annual budget is not reported, I use the reported expenditure on the CCT program.

I supplement data on program characteristics with information on the identity of policymakers, using the Index of Economic Advisers dataset (Kaplan, 2018; Goes and Kaplan, 2024). The Index of Economic Advisers is a dataset of the educational background and training of economic advisors in Latin America and the Caribbean from 1989 to 2022. This gives me a measure of the subject, the level, and the country of education of economic ministers and Central Bank governers for countries in my sample.

I.2.3 Context and descriptives

The final dataset includes 128 program evaluations of the casual impact of CCTs in Latin America and the Caribbean on poverty related outcomes, mapped to annual spending on the evaluated programs. In this section, I present descriptive facts on program evaluations and spending for my sample.

Figure I.1 illustrates the active CCT programs and number of aggregate program eval-

uations for countries in my study. CCT programs in the region are repeatedly evaluated over time. Over the time period 2000 to 2015, there are 31 evaluated CCTs across 17 countries. As seen in Figure I.1b, while Mexico's PROGRESA/ Oportunidades is by far the most heavily studied program, evaluations are common and widespread. The median country has had seven causal evaluations on the impact of CCTs on poverty-related outcomes.

The evaluated CCT programs are large, institutionalised social protection programs with the explicit aim of poverty reduction. Mean spending on CCTs in 2015 was 1,500 million USD, representing 0.29% of GDP in these countries and 17% of the total spending on social protection. Moreover, CCT spending varies annually within programs. Over the 15 year period, the median annual spending increase on programs was 8%, with 35% of program-year observations experiencing decreases in year-on-year spending; and 11% of program-year observations experiencing a more than doubling of spending.

Policymakers over this time period are highly trained and technocratic. Using the Index of Economic Advisors (Kaplan, 2018; Goes and Kaplan, 2024), I find that 60% of finance ministers in the region hold PhDs in 2000s, with the majority of graduate degrees being in Economics (see Figure A3).

What are the origins of the program evaluations? Evaluations are highly embedded in government, suggesting that policymakers are likely to be aware of evidence base (table I.2). 65 of the 128 evaluations are institutional collaborations, wherein the author has a working relationship with the implementing government or international institution. A further 55 evaluations are explicitly demanded by government agencies or international institutions through contracting or funding relationships. 70 are independent evaluations, that are both demanded and evaluated by independent researchers.

The 128 program evaluations in my sample represent 468 treatment effect estimates of the causal impact of CCTs on poverty-related outcomes (table I.1). The size of the treatment effects varies across studies, but most countries in my sample have experienced both a positive, and a negative evaluation result. In particular, almost all countries in

the region have had program evaluations with positive and significant findings on the poverty impact of CCTs (panel a, figure I.3). And moreover, almost all countries have had program evaluations with negative and significant findings on the poverty impact of CCTs (panel b, figure I.3).

I.3 Conceptual framework & method

Given the variation in program spending and evaluation outcomes over time and across countries, it is unclear ex ante which patterns in CCT spending would be consistent with evidence-use. In this section, I therefore begin with a conceptual framework of policy spending under uncertainty (section I.3.1). This framework guides the empirical method, which is described in sections I.3.2 and I.3.3.

I.3.1 Conceptual framework

I present a model of policy spending under uncertainty, based on the model in Buera et al. (2011), who study a policymaker's decision to implement market-oriented policies using their own and neighbours' past experiences. I adapt the learning environment to incorporate policymaking using information signals from program evaluations.

Assume that the policymaker derives utility from minimising the sum of poverty, Y_{it}, and political and social costs, K_{it}, subject to their beliefs of how poverty changes over time.[13] Policymakers choose θ_{it}, an indicator variable for whether or not to increase spending on a CCT program, to maximise their expected utility.[14]

The optimisation problem is thus summarised as follows:

$$\begin{aligned} \min_{\theta_{it}} \quad & E_{it-1}[\log Y_{it} + \theta_{it} K_{it}] \quad \text{s.t.} \\ & y_{it} = \gamma_i \theta_{it} + \varepsilon_{it} \quad \textit{(perceived DGP)} \end{aligned} \tag{I.1}$$

[13] The aim of minimising poverty is consistent with the CCT programs in my sample, all of which have the stated aim of reducing poverty.

[14] I focus on a binary decision for simplicity, but the results of the model hold under a continuous spending variable.

where Y_{it} is the poverty headcount in country i and period t, y_{it} is the implied rate of poverty reduction from observed data, K_{it} is the social and political cost of policy θ_{it} and $\varepsilon_t \equiv [\varepsilon_{1t}, ..., \varepsilon_{nt}]' \sim N(0, \Sigma_\varepsilon)$ is a normally distributed random shock that is correlated with θ_{it} (i.e. $cov(\varepsilon_{it}, \theta_{it} \neq 0)$). The cost, K_{it} is observed to the policymaker (not to the econometrician), but the causal impact of spending on poverty reduction, γ_i, is imperfectly observed.

Timing: In period $t-1$, the policymaker observes signals on the effectiveness of their past policy decisions on the change in poverty, y_{it}. They use this information to update their beliefs of γ_i, the effectiveness of policy spending on poverty reduction. At the beginning of period t, the policymaker then observes the realisation of the political and social cost, K_{it}. Given their beliefs of γ_i, they decide whether or not to increase spending.

The optimal policy decision is therefore given by:

$$\theta_{it}^* = 1[E_{it-1}(\gamma_i) > K_{it}] \tag{I.2}$$

where $E_{it-1}(\gamma_i) = \tilde{\gamma}_{it-1}$ is the policymaker's belief on the effectiveness of increasing cash transfer spending for poverty reduction, as assessed at the end of period $t-1$. That is, policymakers choose to increase spending on a program if the perceived benefit of increasing cash transfer spending is greater than the political and social costs of doing so.

Learning environment: Within this framework, program evaluations can influence policy spending through providing information on γ_i, the impact of spending on poverty reduction. A policymaker is using evidence to make policy decisions if they form evidence-based beliefs – that is, beliefs consistent with evidence from program evaluations. I define evidence-based beliefs as the following:

$$E_{it-1}(\gamma_i) = f(\hat{\boldsymbol{\mu}}_{it-1}) \tag{I.3}$$

where $\hat{\boldsymbol{\mu}}_{it-1}$ is a vector of all evaluation outcomes from program evaluations on country i available before or including in year $t-1$; and $f(.)$ is an increasing function of $\hat{\boldsymbol{\mu}}_{it-1}$[15]. Equation I.3 reflects the idea that outcomes measured in program evaluations tend to be proxy measures or intermediate outcomes of the program objective, i.e. poverty reduction.

Combining equations I.2 and I.3, this implies that there would be a positive relationship between policy spending and evaluation outcomes if the following three conditions hold: (1) policymakers use evidence to update beliefs on the impact of spending, γ_i; (2) evidence is a good signal of γ_i, such that there is a strong mapping between $\hat{\mu}_{it-1}$ and γ_i; and (3) the costs (K_{it}) to increasing spending are moderate, such that there exists periods in which the expected benefit of increasing spending is greater than the corresponding cost, and vice versa, there are periods in which the expected benefit of increasing spending is lower than the corresponding cost [16]. The relationship between evaluation outcomes and policy spending therefore depends not only on the evidence-base, but also on the interaction between features of evidence and costs to policy change.

This basic setup makes explicit the benefits of and barriers to evidence-based policy spending. In a world of uncertainty and limited resources, program evaluations can provide a signal of the causal impact of the program on desired outcomes. Evidence therefore has the potential to increase the efficiency of policy spending by helping policymakers decide which policies to scale up or scale down.

At the same time, there are many reasons why policymakers may not change spending in line with the evidence-base. First, even if policymakers are inclined to use evidence, program evaluations are imperfect because they do not necessarily measure the causal impact of policies on outcomes and policy decisions that are most relevant to the policymaker. As a result, evidence is not always a good signal of γ_i. Second, policymakers may not learn or update their beliefs in a way that is consistent with the evidence [17].

[15] That is, poverty-related outcomes that are measured in program evaluations are positively correlated with actual poverty reduction.

[16] Intuitively, this means that the social and political costs are not too high or too low, such that condition I.3 always or never holds.

[17] This is the focus of much of the existing literature on research use in policy, which focuses on measur-

Third, even if policymakers learn from the evidence, the expected benefits of changing spending will not necessarily overcome the costs to evidence-use. The costs of increasing policy spending vary by context, and are likely to depend on factors such as the electoral cycle, political competition, and public sentiment. These costs may also interact with features of the evidence base. At the extreme, evaluations that measure politically salient outcomes or that can be attributed to the policymaker may be associated with low or even negative political costs.[18]

Given the ambiguous theoretical relationship, it is therefore key to establish the baseline relationship between causal estimates of impact and spending empirically in the data. The empirical relationship of interest can be summarised as the following:

$$\Delta log(spend_{it}) = \alpha + \beta f(\hat{\boldsymbol{\mu}}_{it-1}) + \varepsilon_{it} \tag{I.4}$$

where $\hat{\boldsymbol{\mu}}_{it-1}$ is the vector of evaluation outcomes of a CCT program in country i and in year t, and $\Delta log(spend_{it})$ is the change in log spending on the evaluated CCT program in year t relative to year $t-1$. Under assumptions outlined above, $\beta > 0$ is consistent with models of evidence-based policy spending.

The main empirical challenge of estimating equation I.4 is in estimating $f(\hat{\boldsymbol{\mu}}_{it})$, the perceived causal impact of a CCT program based on a given evaluation. While $f(\hat{\boldsymbol{\mu}}_{it})$ is known to the policymaker, it is unobserved by the econometrician. This is due to two main reasons:

- The econometrician cannot observe $\hat{\boldsymbol{\mu}}_{it}$, the **information set** that is relevant to the policymakers at each point in time; and
- Even if the information set were known, the econometrician cannot observe how policymakers aggregate information both within and across studies, $f(.)$. In other words, the **aggregation method** is also unobserved.

ing belief updating and willingness to pay for evidence (Vivalt and Coville, 2023; Banuri et al., 2017; Hjort et al., 2021).

[18] For instance, in settings with an informed electorate, voters can discipline politicians by threatening to replace incumbents in elections.

I thus estimate equation I.4 by constructing estimates of $f(\hat{\boldsymbol{\mu}}_{it})$ – summary metrics of impact – across different information sets and aggregation methods, which together, capture different patterns of evidence-based policy spending. First, I consider the marginal impact of individual evaluations, summarised by aggregated metrics of information from each individual study. Second, I consider the impact of aggregate bodies of evidence, summarised by the posterior mean of aggregate country-level findings from a Bayesian hierarchical model.

I outline the aggregation methods used for individual papers and for each country's evidence-base in sections I.3.2 and I.3.3.

I.3.2 Aggregating results from individual evaluations

I begin by exploring the relationship between individual evaluations and subsequent spending. I consider the relationship between estimated treatment effects from program evaluations of program i, first made available in year $t-1$, and subsequent changes in spending on the same program in t. In particular, I estimate the following linear relationship:

$$\Delta log(\text{spend})_{it} = \alpha + \beta \hat{\mu}_{ist-1} + \varepsilon_{it} \tag{I.5}$$

where $\hat{\mu}_{ist-1}$ is the aggregated paper-level finding from a program evaluation s that evaluates the impact of a CCT program in country i, with $t-1$ being the year that results from the evaluation were first made available. Standard errors are clustered at the country level.

Importantly, each individual program evaluation conveys a multitude of information that is likely to be associated with program impact. This includes both quantitative measures, such as the treatment effect, statistical significance, and the standard error; as well as qualitative information, such as descriptive facts, and the language used to describe the evaluation results.

I therefore consider three aggregations or measures of evaluation outcomes from each individual evaluation. Each of these aggregation methods provides a different estimate

of the evaluation outcome, $\hat{\mu}_{ist-1}$, from each study.

Reported Treatment Effects: I begin by estimating $\hat{\mu}_{ist}$ using paper-level aggregations from reported treatment effects. Program evaluations often include results from multiple econometric specifications on a range of outcomes and populations of interest. I therefore aggregate paper-level results across four metrics: the maximum magnitude, the maximum significance, the mean magnitude, and the mean statistical significance of headline results. I capture the magnitude of the causal impact of CCTs by the effect size, calculated as the estimated treatment effect divided by the standard deviation [19]; and the statistical significance of research results by the test-statistic, calculated as the estimated treatment effect divided by the standard error.

An illustrating example: Consider the Galiani and McEwan (2013) evaluation of the Programa de Asignación (PRAF), a CCT program in Honduras. The authors find that PRAF causally reduced the prevalence of child labour by 3 percentage points (se = 0.011, effect size = 0.017) and increased the probability of children attending school by 8 percentage points (se= 0.023, effect size = 0.020). I thus consider four study-level summary statistics to capture the range of different potential signals from the same program evaluation (see table I.3): 0.017, capturing the maximum magnitude of headline results; 3.48, capturing the maximum significance; 0.020 capturing the mean magnitude; and 3.10, capturing the mean significance of headline results.

Evaluation results, relative to the existing knowledge base: As documented in section I.2, CCTs are often evaluated repeatedly over time. The median country in my sample is evaluated seven times, with almost every country having had over three evaluations from 2000 to 2015. Program evaluations on CCTs therefore contribute to an existing stock of knowledge on the impact of cash transfers on poverty related outcomes. Hence, rather than responding to *reported* treatment effects from program evaluations, policymakers may be more responsive to findings that they find 'surprising', relative to their

[19] Most papers do not report the standard deviation of the control group. This means that in practice I compute the within-group standard deviation using the standard error of the difference in means, from the estimated treatment effect. This gives me an estimate of the average standard deviation of the treatment and control groups, and is comparable to the standard deviation of the control group under the assumption that the two groups have the same variance.

existing prior beliefs.

To explore responses to *surprises* from the causal studies, I summarise paper-level findings as:

$$\hat{\mu}_{ist} = \tau_{ist} - \hat{v}_{it} \tag{I.6}$$

where $\hat{v}_{it}$ is a measure of the prior beliefs on the effectiveness of cash transfers based on the existing stock of findings available up to year t, and τ_{its} is the aggregated paper-level treatment effect from paper s, country i, and available in time t. $\tau_{ist} - \hat{v}_{it}$ is therefore a measure of how 'surprising' a paper is, relative to the existing evidence base. $\tau_{ist} - \hat{v}_{it} > 0$ means that the CCT is performing better than would be expected; whereas $\tau_{ist} - \hat{v}_{it} < 0$ means that the CCT is underperforming, relative to expectations.

To estimate $\hat{v}_{it}$, I assume that policymakers form prior beliefs based on the existing evidence base, in a manner that is consistent with fixed effects. That is, I estimate $\hat{v}_{it}$ as a precision weighted mean the findings from the cumulative evidence available at time t.

The implied prior belief based on the cumulative stock of knowledge is given by:

$$\hat{v}_{it} = \frac{\Sigma w_{is} \tau_{ist}}{\Sigma w_{is}}$$

$\forall s \in t$ where:

$$w_{is} = \begin{cases} \frac{1}{\sigma_s^2}, & \text{if } i = j \\ \lambda \times \frac{1}{\sigma_s^2} & \text{if } i \neq j \end{cases}$$

where σ_s^2 is the precision of study s, and $\lambda \in [0, 1]$ is the weight placed on research published in other countries.

Critically, λ allows for some flexibility in assumptions on the weight that policymakers place on research results from other countries. When $\lambda = 0$, the policymaker believes there is zero external validity, and therefore only forms expectations based on prior research from their own country. At the other extreme, when $\lambda = 1$, the policymaker believes there is perfect external validity, and places equal weight on research from all countries. I construct estimates of $\tau_{ist} - \hat{v}_{it}$ across values of $\lambda \in [0, 1]$, using the mean test

statistic and the mean effect size of each paper.

Framing of research results: Beyond the magnitude and significance of treatment effects, politicians may instead be responsive to how research results are described and communicated. In describing study findings, researchers convey evaluation results through their choice of language. This, in turn, can affect the beliefs and decision-making of consumers of research. For instance, Dylong and Koenings (2023) find that the framing of expert GDP forecasts as positive news, relative to existing growth trajectories increases policy support. In the presence of time and cognitive constraints, policymakers may rely on the author's language and framing of the research findings to make policy conclusions.[20]

To explore the importance of the framing of research results, I summarise $\hat{\mu}_{ist}$ by the abstract sentiment score. I use the Hu and Liu (2004) lexicon to classify each word of the paper abstract into positive, neutral, or negative sentiment phrases. The abstract sentiment for each paper is defined as:

$$\hat{\mu}_{its} = \text{Abstract sentiment} = \frac{\text{N positive} - \text{N negative}}{\text{Total word count}} \tag{I.7}$$

Thus, a positive sentiment score corresponds to a more positively framed abstract – wherein the author(s) have framed the paper findings as more 'positive'.

I.3.3 Aggregating a country's evidence-base

What if policymakers are responding to the cumulative body of evidence? First, there is growing evidence on the prevalence of site-selection bias (Allcott, 2015) and of limited external validity in the presence of stochastic shocks (Rosenzweig and Udry, 2020), both of which limit the potential for learning from individual program evaluations. Second, changing spending in line with the evidence may also take time, due to institutional and political costs to policy change. As a result, evidence-based policy spending may be

[20] Relatedly, Cavallo et al. (2017) find that individuals place weight on less reliable sources of information when forming inflation expectations, even when more reliable information on inflation forecasts are available.

reflected through patterns in aggregate spending and the cumulative evidence-base over time.

I therefore consider the relationship between the aggregate evidence base and spending, as follows:

$$log(spend)_i = \alpha + \beta\hat{\mu}_i + \varepsilon_i \tag{I.8}$$

where $\hat{\mu}_i$ is the estimated posterior mean of findings from all CCTs evaluations conducted on country i by 2015; and $spend_i$ is the spending on CCT programs in country i in 2015.

I estimate $\hat{\mu}_i$, the aggregated measure of cumulative findings from a country's evidence-base, using a two-stage Bayesian hierarchical model. The Bayesian hierarchical model tackles challenges of aggregation by jointly estimating the heterogeneity in treatment effects that arises from sampling variation, due to noise at the study-level, versus genuine heterogeneity, due to true variation in treatment effects. The posterior mean from the hierarchical model therefore gives an estimate of the true average that optimally shrinks the population mean towards more informative studies. Bayesian hierarchical models are common in the meta-science literature, and is increasingly used in economics (e.g. Meager, 2019; Bandiera et al., 2022).

My model consists of two-stages, and extends the canonical Rubin's eight schools model (Rubin, 1981). In the first stage of the estimation, I aggregate the treatment effects within each evaluation to obtain an estimate of the posterior mean for each program evaluation. This gives an estimate of the evaluation outcome at the study-level. In the second stage of the estimation, I use the posterior estimates of study-level findings from the first stage to estimate a country-level posterior mean of the cumulative evidence base. This gives an estimate of the aggregate impact of CCTs in a country.

First stage. Let $\hat{\tau}_{kji}$ be the reported treatment effect k from evaluation j, which studies the causal impact of CCTs in country i. $\hat{se}^2_{kji}$ is the associated standard error of the estimated treatment effect. Each evaluation has between one to six main reported treatment effects (headline results). For each evaluation j, I estimate the posterior mean of the

evaluation, $\hat{\tau}_{ji}$, as:

$$\hat{\tau}_{kji} \sim N(\tau_{kji}, \hat{se^2}_{kji}), \qquad k = 1...K$$
$$\tau_{kji} \sim N(\tau_{ji}, se^2{}_{ji})$$

Second stage. Using the posterior mean of the evaluation treatment effect and standard error, $\hat{\tau}_{ji}$ and $\hat{se^2}_{ji}$ from the first stage, I then estimate a country-level posterior mean using the following:

$$\hat{\tau}_{ji} \sim N(\tau_{ji}, \hat{se^2}_{ji}), \qquad j = 1...J$$
$$\tau_{ji} \sim N(\tau_i, \sigma^2_\tau)$$

The estimate of τ_i from the second stage gives me an estimate of the posterior mean of the country-level treatment effect, based on all program evaluations of CCTs conducted in country i, between 2000 to 2015.

To estimate the model, I use weakly informative priors on the hyperparameters, which underlies the assumption that absent the evidence, policymakers believe that the program has zero impact. The main assumption of the model is that of exchangeability between effect estimates. In practice, this implies that absent seeing the study estimates, there should be no reason to believe that the average impact of cash transfers is greater in one study versus another. I estimate the posterior distribution of the model via simulation, using Hamiltonian Monte Carlo methods (HMC).

I.4 Individual evaluations & spending

I.4.1 Reported treatment effects

I begin by aggregating findings within each evaluation using the mean of the t-statistic of headline results. The t-statistic – calculated as the treatment effect divided by the standard error – captures the statistical significance of findings and is the most consistently reported and comparable statistic across all program evaluations in my sample. In a two-sided test, a t-statistic that is less than or equal to -1.65 represents a negative treatment effect that is statistically distinguishable from zero at 10%; whereas a test statistic that is greater than or equal to 1.65 represents a positive treatment effect that is statistically significant at 10%. [21]

In figure I.4, I plot the baseline relationship between the mean significance of each paper, and subsequent spending on the same program. More significant evaluation-level findings do not correspond with larger increases in spending.

This zero correlation is not driven by choices in the aggregation or measure of reported treatment effects. In figure I.5, I plot the relationship between causal estimates of program impact and spending across four summary measures of headline results: the maximum magnitude, the mean magnitude, the maximum statistical significance, and the mean statistical significance. Across all four ways of summarising paper-level findings, I find there is no systematic relationship between estimates of impact and subsequent spending on the same program.

I consider the responsiveness in spending to paper-level findings using only within-country or within-year variation. As seen in table I.4, the null relationship is not driven by fixed, unobserved country or time characteristics that are correlated with evaluation findings and spending decisions.

[21] For comparability, I adjust treatment effects such across all outcome categories so that a positive treatment effect or test statistic is interpreted as a welfare improving outcome; and a negative treatment effect or test statistic can be interpreted as a 'bad' outcome.

The estimated null relationship is small in magnitude and relatively precise. A coefficient of 0.058 on the mean treatment effect implies that moving from a mean t-statistic of 0 to 1.96 would correspond with a $1.65m increase spending. This accounts for less than 1% of the mean annual change in spending, and less than 0.1% of the mean annual spending on CCT programs across this time period. At the upper bound of the 95% confidence interval, the estimated coefficient would account for less than 5% of the mean annual change in spending, and less than 0.5% of the mean annual spending.

To what extent are these findings driven by policymaker awareness of evaluations? The policymaker's consumption of evidence is unobserved. However, I can proxy for policymaker awareness using information on the source of the program evaluation, and government-author relationships. In figure I.6, I plot the estimated relationship between evaluation outcomes and subsequent CCT spending, by government demand and relationships. First, I consider the subset of studies that are conducted by authors that have a relationship with government (*Author-gov link*). These studies could be associated with higher take-up, both because policymakers are more likely to be aware of the evaluation results, and because the authors are more likely to measure outcomes that are pertinent to the policy environment. For instance, Bonargent (2024) finds that projects developed in partnership with policymakers are up to 20 percentage points more likely to result in policy change. I find that the estimated magnitude is larger for this subset of studies, but it is statistically indistinguishable from zero. Second, I consider the subset of evaluations that are explicitly demanded by government or international institutions (*Institutional evaluations*). Again, I find a null relationship between the evaluation outcomes and changes in spending. This suggests that the zero relationship is not driven by lack of policy awareness.

How do these results relate to organisational or political costs? Even if policymakers are aware of evaluation outcomes, and use evidence to update their beliefs, this would only translate to changes in spending if the perceived benefits of doing so outweigh the costs to policy change. Evaluation results made available in years with lower political or organisational costs to policy change may therefore be associated with higher respon-

siveness to treatment effects.

To examine the role of organisational costs, I consider different assumptions around the timing of spending increases, relative to when research results are made available. Policy spending may take time to implement, in which changes in CCT spending would only be reflected through longer time lags. The linear relationship between treatment effects and changes in spending up to three years after the release of evaluation results are statistically indistinguishable from zero at the 5% level across all four measures of treatment effects (tables I.6 and I.7).

I explore the role of political costs in figure I.7, by considering the association between treatment effects and spending across different baseline political conditions at the year in which the evaluation results were released. The political returns of increasing CCT spending is likely to differ in election versus non election years. I explore these patterns in figure I.7, by considering responsiveness in election versus non election years. I find no evidence of differential responsiveness across election versus non election years.

I find a robust zero association between spending and reported treatment effects across various evaluation-level aggregations of headline results. One explanation for this may be that policymakers have strong priors on the size of the treatment effects, such that they correctly anticipate the program evaluation results. In this case, the signal from each evaluation depends on how surprising the finding is, relative to the existing evidence base. I therefore quantify the surprise from each individual evaluation in section I.4.2.

I.4.2 Quantifying the surprises

In this section, I estimate the size of evaluation-level findings relative to existing potential beliefs from the cumulative evidence base – the 'surprise' from each program evaluation. I estimate the size of each evaluation-level finding relative to the existing prior beliefs across different assumptions on λ, the perceived external validity of studies from other countries.

Assumptions on λ are central to whether the same research finding is interpreted as

a positive or negative shock. I Illustrate this in figure I.8, where the solid lines indicate estimates of evidence-based priors given existing evidence, and the dots represent the mean headline result from each evaluation first made available in each year. In panel a), figure I.8. I plot the evidence-based prior belief when there is zero weight placed on research from other countries ($\lambda = 0$). Here, the program evaluation highlighted in orange is perceived as a negative shock (bad news), since the evaluation outcome is lower compared to existing priors. In contrast, when beliefs are formed by placing equal weight on all papers available in the region ($\lambda = 1$), the same evaluation is perceived as a positive information shock (good news). Hence, the same evaluation can be perceived as a positive or negative shock, depending on policymaker beliefs on the external validity of evaluations from other countries (panel b, figure I.8).

I therefore estimate the relationship between evaluation surprises and changes in spending across different assumptions of λ, from 0 to 1. In figure I.9 I plot the estimated $\hat{\beta}$ and 95% confidence intervals from a linear regression of equation I.6. Across all assumptions of external validity, more surprising findings do not correspond with larger changes in spending.

Are there asymmetric responses in spending, with respect to positive versus negative findings? Negative findings that underperform relative to expectations may hold greater weight than positive findings because they suggest that programs are not working as well as anticipated. However, withdrawing spending from a CCT program may be costly, especially given the political saliency of CCTs. Moreover, findings from belief-elicitation experiments suggest that policymakers exhibit asymmetric optimism and update their beliefs more in response to positive research results (Vivalt and Coville, 2023). I examine evidence for both of these channels, by considering the relationship between subsets of evaluation results that are more positive and more negative, relative to the existing evidence-base (figure I.10a, figure I.10b). I find a consistent zero relationship for both positive and negative surprises.

I.4.3 Framing of research results

While I have thus far focused on the treatment effects of headline findings, authors can also communicate the strength of evaluation outcomes through the language they use to describe the research results. I therefore move beyond aggregations of reported headline results, to consider patterns of spending responsiveness to the framing of research results.

As outlined in section I.3.2, I estimate the framing of research results by the sentiment score in the abstract (defined by equation I.7). In general, the abstract sentiment score of evaluations tends to be positive, reflecting the idea that authors are inclined to use more positive than negative language to describe research findings. In figure I.11, I plot the relationship between the mean significance of headline results and the abstract sentiment score in each paper. 29 papers with negative or null results are still positively framed.

In figure I.12, I plot the relationship between the abstract sentiment score and changes in spending on the same cash transfer program. I find that positively framed results are not systematically correlated with larger changes in spending. Thus, the results cannot be explained by higher policy responses to more optimistic or positively framed evaluation results.

I.5 Cumulative evidence & spending

As seen in section I.4, I find no evidence that policymakers adjust their spending in response to individual evaluations. Nonetheless these patterns can be consistent with evidence-based policy spending if, instead of responding to individual papers, policymakers learn and adjust their spending over time in line with the cumulative evidence base. In this case, evidence-based policy spending would be observed as higher spending in countries with programs that have been shown to be more impactful.

Using the two-stage Bayesian hierarchical model outlined in section I.3.3, I estimate the posterior mean of each country's findings given the entire body of evidence. In figure I.13, I plot the posterior mean of aggregate results for each country from the second stage

of the hierarchical model against the log of cash transfer spending in 2015, the final year of my study period [22]. I find that there is no relationship between cumulative findings at the country level and CCT spending.

This result holds when considering the relationship with spending as a share of GDP, and as a share of the total social protection budget in 2015 (see Section A.2 of the Appendix). It also holds for cruder aggregations of the evidence-base, such as the arithmetic mean of treatment effects.

The absence of empirical evidence for cumulative learning could be explained by program evaluations not being generalizable to the study population. The Bayesian Hierarchical framework provides of natural measure of this through the pooling metric defined in Gelman and Pardoe (2006). I estimate the summary pooling factor for each country as follows:

$$\gamma_i = 1 - \frac{\sigma^2_{\tau i}}{\sigma^2_{\tau i} + E_j(se^2_{ji})} \tag{I.9}$$

γ_i is bounded between 0 and 1, and gives an estimate of the proportion of the total variation that can be explained by variation in the study. $\gamma_i > 0.5$ indicates a reasonable amount of pooling, suggesting that there is more information at the population level than at the study level. This implies that studies are more likely to be estimating a common mean – and hence, is suggestive of higher external validity.

Figure II.3 illustrates the estimated γ_i for all countries with more than three studies. As seen from the figure, almost all countries have a pooling factor greater than 0.6. This implies that there is considerable amount of pooling across studies, and suggests that external validity is relatively high. Thus, program evaluations are likely to be informative about the populations of interest.

[22] I examine the cross-country relationship between spending and aggregate findings in a single year (2015). This is because spending on CCTs is highly autocorrelated and by construction, the cumulative treatment effect for each country is also highly autocorrelated across time.

I.6 Discussion

Overall, I find a robust and relatively precise zero relationship between policy spending and causal estimates of impact across paper-level aggregations (section I.4) and country-level aggregations (section I.5) of the evidence base. The average zero relationship suggests that either policymakers do not adjust their spending in response to causal estimates of impact, or there is a complex relationship that directly offsets any changes made, resulting in a reliable zero correlation. Given program evaluations of CCTs are highly embedded in government, this result seems unlikely to be driven by lack of policy awareness, but is suggestive of the presence of inefficiencies or costs to policy change.

What do these findings tell us about alternative models of evidence-based policy spending? One alternative model would be the use of evidence on comparative policies for relative spending decisions. If comparative policies to CCTs are consistently shown to have higher returns than CCTs, then evidence-use would be observed by a re-allocation of spending away from CCT spending. This type of evidence-use seems unlikely to be driving the results for two reasons. First, comparative large-scale policies for poverty alleviation are not evaluated as heavily or systematically as CCTs. Illustratively, the Development Evidence portal records 205 published impact evaluations on social protection policies in LAC countries between 2000 to 2015 (International Initiative for Impact Evaluation, 3ie, 2024). The vast majority of these studies (135 studies out of 205) examine the causal impact of Conditional Cash Transfers. Following CCTs, the most frequently evaluated social protection programs are Unconditional Cash transfers (23 studies); and food transfers (12 studies)[23]. Therefore, policymakers are unlikely to have alternative rigorous evidence on comparative policies. Second, the null result holds when considering the relationship between evaluation outcomes and CCT spending, as a percentage of social protection expenditure. This suggests that the zero relationship is not driven by policymakers reallocating spending to alternative social protection policies, in periods

[23]Relatedly, very few program evaluations of CCTs study the causal impact of CCTs, compared to the causal impact of an alternative policy. Therefore, the evaluations in my study all focus on the impact of being a CCT recipient compared to a counterfactual outcome of being a non-recipient.

where CCT evaluation outcomes are higher.

Another challenge in interpreting the average zero relationship is the fact that the policymaker's objective function is unobserved. If, for instance, the policymaker is not aiming to minimise poverty (as assumed in section I.3.1), but rather, aiming to achieve a target poverty rate, the observed relationship between policy spending and program evaluations would be zero, even in the presence of evidence-based policy spending. This would not be discernable from the data.[24]

Despite these limitations, the zero average relationship can shed light on other objective functions that are common to discussions around evidence-use. For instance, Kremer et al. (2021) estimate a social benefit-to-cost ratio of development innovation, which underlies a model in which policymakers would be maximising on the cost-effectiveness of policies. If policymakers are maximising on cost-effectiveness, evidence-based policy making would still translate to a positive average relationship between evaluation outcomes and spending unless there is an inverse relationship between program impact and costs, such that programs are more costly when they are less impactful. In practice, it seems unlikely that policymakers are maximising on cost-effectiveness, as systematic reports of cost-effectiveness are uncommon, and particularly difficult to estimate in the context of CCTs (Evans and Popova, 2016).

Another common narrative of evidence-use in policy is related to the exogeneity of evaluations. 32 evaluations in my sample are explicitly demanded by implementing governments. This may bring concerns of potential 'impact buying' wherein policymakers pay for research results to justify desired future spending changes. If this were the case, the partial relationship between spending and program evaluations would likely be an upper bound of the true causal impact, since policymakers would be more likely to commission evaluation results that are positively correlated with their desired changes in policy spending.

I provide two pieces of evidence which suggest that this form of impact buying is not

[24] This control function objective does not match the documented stated objectives of the CCT programs, however.

of first order concern. First, I find that government demanded evaluations tend to be set up from the inception of the program. The evaluation of PROGRESA/Oportunidades established a tradition of evaluating CCTs from the onset of program design (Rawlings and Rubio, 2005). Therefore, the timing of evaluations suggests that there is limited presence of impact buying by governments. Second, I consider the relationship between spending and paper level findings for the subset of independent evaluations, that are both demanded and evaluated by independent institutions. Within this subset of evaluations, I find that there is no association between paper level findings and spending (figure I.6).

While the focus of my study is the intensive margin of spending, program evaluations could also affect other policy decisions. Evaluations could have increased program survival by making it less likely for countries to end programs. As negative evaluations do not correspond with decreases in spending, my intensive margin results suggest that evaluation outcomes are not predictive of when programs end. Nonetheless, almost all countries in my sample have had at least one positive and significant evaluation between 2000-2015, and have an active CCT program in place in 2015. Therefore, my findings could be consistent with the narrative that the existence of any program evaluation contributed to the longevity of CCT programs in the region. Given that CCTs are highly salient policy instruments, this interpretation would be suggestive of the political value of positive evaluation outcomes. The existence of any positive program evaluation could act as a 'stamp of approval' , helping policymakers build longer-run political support to sustain CCT programs.

Furthermore, program evaluations could have had an indirect impact on policy decisions, by influencing other sources of policy relevant information. The importance of this channel depends on the extent to which findings from evaluations are transmitted to other information sources. Given I do not observe the full information set available to policymakers, whether this is the case is unknown. Lastly, the evaluation of CCTs could have led to the spread of CCT programs worldwide, by building consensus around the effectiveness of CCTs for poverty alleviation. I leave both of these channels for future research.

I.7 Do features of evidence matter?

In section I.3.1, I implicitly assume that all program evaluations are relevant to policymakers aiming to learn about the impact of their programs. If, however, evidence-use is costly, policymakers may be rationally selective on the subset of evaluations that they use to form decisions on policy spending. That is, they may limit the information set (μ_{it}) to subsets of evaluations that are more relevant for policy decisions. Importantly, the choice of policy relevant evaluations may further interact with political and practical costs to policy change, K_{it}, as evaluations with certain characteristics may be associated with lower costs to evidence-use.

In this section, I consider heterogeneous responses in spending along three dimensions of evidence characteristics that are often associated with greater suitability for policy decisions.

1. More credible evidence, defined as program evaluations that are more internally valid, or associated with higher academic quality;

2. More generalizable evidence, defined as program evaluations that are more externally valid or relevant to the population of interest;

3. More actionable evidence, defined as program evaluations that are more timely, or embedded in the policymaker's decision process.

Credible evaluations can be more conducive to learning because they provide higher quality or more reliable estimates of the underlying causal effect of interest. Politicians that place greater weight on the internal validity of studies may be more responsive to studies that use experimental variation to identify the causal effect of interest. There is some evidence that this is the case. For instance, Mehmood et al. (2021) finds that policymakers place greater weight on experimental studies after being trained in causal inference methods. Beyond the methodology of a study, policymakers may also place greater weight on studies that are peer-reviewed and published in top academic journals

if they are seen as more credible.

Even if policymakers do not explicitly value credibility, the costs to evidence-use may still be lower for this subset of evaluations. Results from randomised controlled trials may be easier to implement, as they are often seen as the 'golden standard' of evidence. Similarly, program evaluations that are published in top academic journals may be more difficult to refute. Studies of this type may therefore be more likely to correspond with policy change.

In figure I.15 (table I.8), I plot the association between the mean t-statistic and subsequent spending for subsets of studies, across different measures of credibility.[25] First, I consider selective responsiveness to randomised controlled trials, evaluations that use random variation to identify the causal estimate of interest. I find no evidence of responsiveness to experimental studies. The coefficient estimate for the subset of studies that are RCTs versus non-experimental are similar in magnitude. I then consider selective responsiveness by the academic quality of the program evaluation, using an indicator of whether the evaluation is published in a top 100 academic journal[26]. I find no evidence of selective responsiveness to academic quality.

Beyond credibility, program evaluations differ by how generalizable they are to the population of interest. This is important for policy, because while program evaluations may be internally valid, they may be less informative about the impacts of the program to the broader population of interest. This means that evaluations that are internally valid but not broadly more generalizable are likely to be less useful for policy decisions.

I measure the credibility of each study using the pooling factor from the Bayesian hierarchical model given in equation I.9. A higher pooling factor implies that there is considerable pooling across studies, which suggests that there is a reasonable amount of external validity across studies. I define a study as having a high pooling factor when the pooling factor is greater than 0.6. I also consider more direct measures of generalizability,

[25] Here, and in this section, I focus on the t-statistic as the summary metric for each individual evaluation, as this is the only statistic that is consistently reported across studies.

[26] I use the journal rankings from REPEC to classify whether the program evaluation is from a top academic journal.

by using the population of interest pertaining to the program evaluation. Around half of the evaluations in my sample study the causal impact of CCTs on poverty-related outcomes for only rural or urban sub-populations. I consider the association between spending and evaluation for this subset of studies, versus those that study the causal impact of CCTs for the full population.

As seen in figure I.16 (table I.9), there is zero association between treatment effects and subsequent spending for both high and low pooling studies. Similarly, there is zero association in spending both, across sub-population studies and evaluations that study the treatment effect of the full population.

How actionable and embedded are program evaluations for policymaker decisions? I consider two main dimensions of actionability, as proxied by the outcomes and the timeliness of evaluation.

Results from evaluations may be more actionable for policy decisions if they measure outcomes that are better aligned with the objectives and decisions relevant to the policymaker's decisions. While all program evaluations in my sample study the impact of CCTs on poverty-related outcomes, these outcomes can be further classified into subcategories, including: education, health and nutrition, gender, employment, and savings, investment, and production. Given that CCTs in my sample often explicitly condition on education and health behaviours, evaluations that explicitly study the causal impact of programs on these outcomes may have more actionable implications for policy decisions. Alternatively, other outcome categories, such as employment, could be more closely aligned with economic policy agendas - and hence, have lower associated costs to policy change. In figure I.17 (table I.10), I plot the association between spending and each of the outcome sub-categories and find a consistent zero relationship.

Beyond outcomes, I explore patterns of spending with respect to the timeliness of individual evaluations. In identifying the causal effect of CCTs, program evaluations study the impact of programs at a given point in time – the effect year. For experimental studies, the effect year corresponds to the endline year of data collection. For non-experimental

studies, the effect year corresponds to the year at which the post-treatment outcome is measured in the data.[27]

I measure the timeliness of evaluation as the number of years between the first year of publication, and the effect year. As seen in figure I.18, the timeliness of evaluations varies largely across studies. Program evaluations are made available up to 13 years after the effect year, with the mean study being published 4 years after the study period.

A longer lag between publication and the effect year is likely to correspond with lower actionability. This is because the evaluation outcomes are less likely to be embedded in the current policy environment, especially in the presence of changes in the policy or economic environment over time. Furthermore, in the presence of time-stochastic aggregate shocks, the dynamic returns of the same policy can change over time (Rosenzweig and Udry, 2020). This decreases the external validity of evaluations that study time periods further in the past. I use variation in the timing of evaluation results to consider differential responsiveness to the timeliness of evaluation. I define an indicator variable, *Timely*, equal to 1 when the gap between the first year of publication and the effect year is within the mean of 4 years.

As seen in figure I.19 (table I.11), I find a positive association between spending and the mean t-statistic for more timely studies. The coefficient estimate of 0.01854 (se=0.0057, p=0.008) is positive and significant at 1%. The coefficient estimate implies that moving from a mean t-statistic of 0 to 1.96 is associated with an increase in spending of around 5.4m USD, accounting for around 3% of the average annual increase in spending. This finding is robust to different definitions of timeliness. In figure I.20 I show that the positive association persists for all studies that are released within the mean of 4 years after the effect year.

The importance of time-actionable results is driven by periods in which the political costs to policy change is lower. In figure I.21 (table I.13), I consider how the responsiveness in spending for timely evaluations interacts with changes to the political party in power. If the results of the evaluation can be attributed to the same political party as that

[27] e.g. In a difference-in-differences estimator, this would be the post-treatment period.

in power at the date of publication, there may be greater political will to change policy spending in line with the evidence – and hence, lower (or even negative) political costs to policy change. I find that when the political party in power is unchanged at the effect year and at the year of publication, there is a stronger association between the treatment effect and subsequent changes in spending. This suggests that the actionability of research findings to spending is higher when evaluations are timely, and when political costs of policy implementation are low.

In contrast, the importance of timely studies does not seem to be driven by other characteristics associated with timeliness. Timely papers are more likely to use non-experimental variation and to have an author that works in government, when compared to non-timely papers (table I.12). In figure I.22 (table I.14), I plot the responsiveness within characteristics of timely versus non-timely papers. The patterns suggest that the findings on the timeliness of results are not driven by measurable study characteristics that are common to timely versus non-timely papers. Lastly, I find a zero relationship between evaluation outcomes and changes in spending one year after the effect year, the earliest date at which policymakers could be aware of the evaluation outcomes (see: figure A5 in the appendix). This suggests that the result is not driven by policymakers being aware of timely studies and incorporating study findings prior to the first year of publication.

I.8 Conclusion

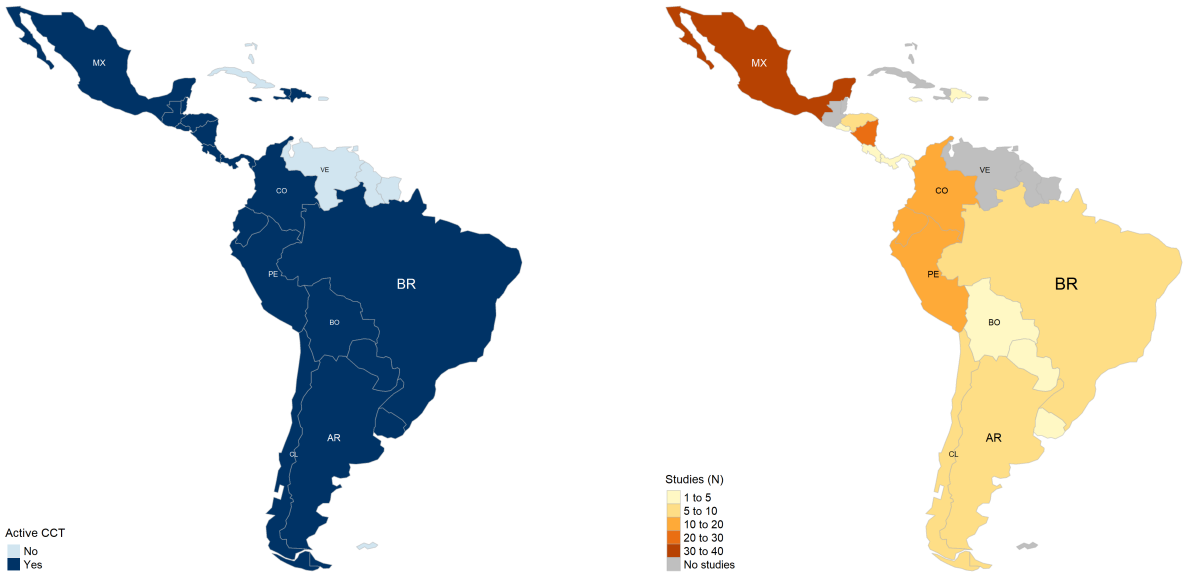
Over the past two decades, there has been a vast increase in the number of program evaluations conducted in academia, government, and organisations. In providing causal estimates of impact, these evaluations can in theory influence policy spending decisions, by helping to channel resources into programs with greater impact. Despite this, there is limited empirical evidence on the relationship between evaluation outcomes and changes in policy spending.

Across 128 program evaluations of Conditional Cash Transfers in Latin America and the Caribbean, I find a robust zero correlation between causal estimates of impact and subsequent policy spending. The only exception is when research results are timely, and when political costs are low. This suggests that the timeliness of evaluation is an overlooked mechanism for increasing the use of evidence in policy. Understanding when research is most impactful, and developing methods to deliver on quick and rigorous evaluations is a valuable avenue for future research and policy.

More broadly, there is considerable scope for increasing the impact of evidence through rigorous empirical analysis on the existing relationship between research and policy. A necessary starting point to this agenda is systematic data collection on the use and engagement with evidence across all stages of the evidence-to-policy pipeline – many of which remain under-explored. Only by understanding this relationship, can we better design research to reach the full potential of evidence-based policymaking.

I.9 Tables and figures

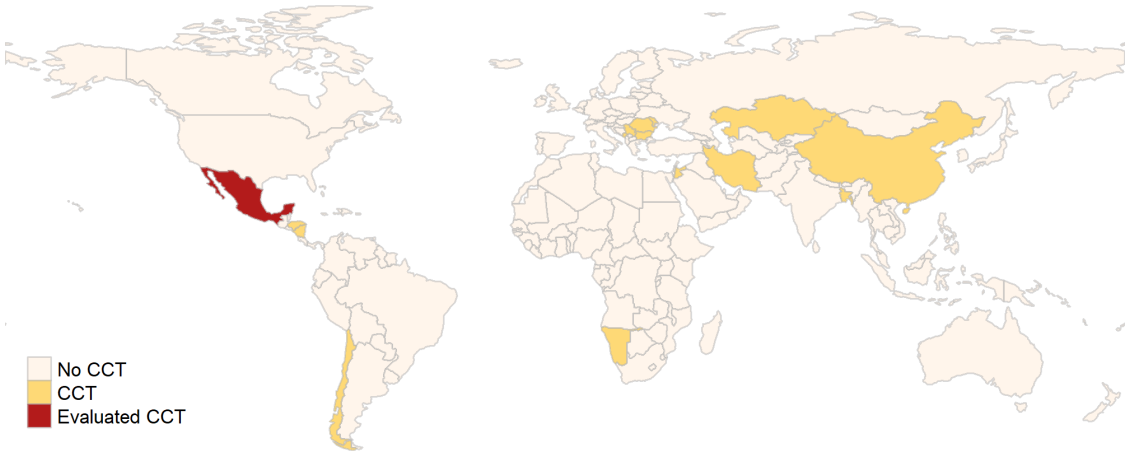
I.9.1 Figures



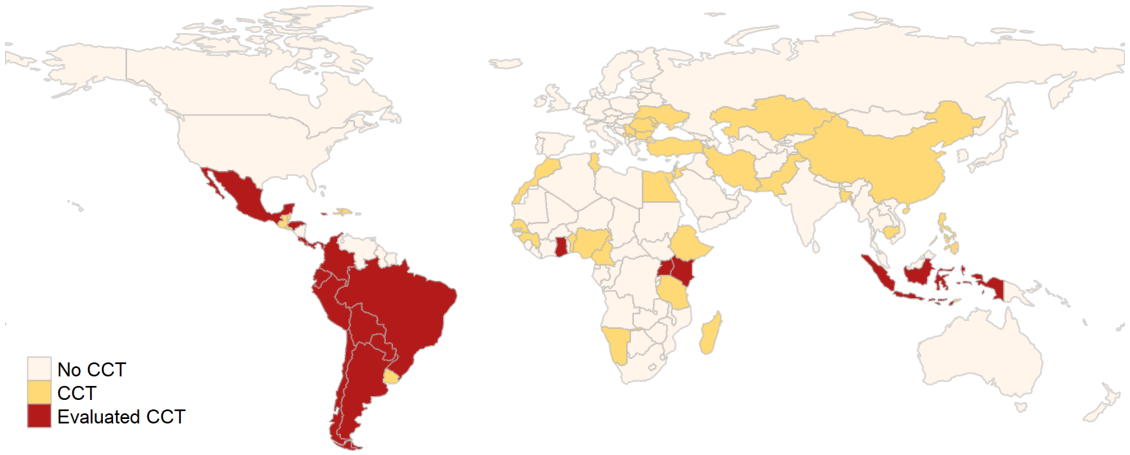
(A) Active Conditional Cash Transfers (B) N. program evaluations by country

FIGURE I.1: Active cash transfers and cumulative program evaluations in 2015

Notes: Active CCTs and number of aggregate program evaluations on CCTs by country in Latin America and the Caribbean in 2015.



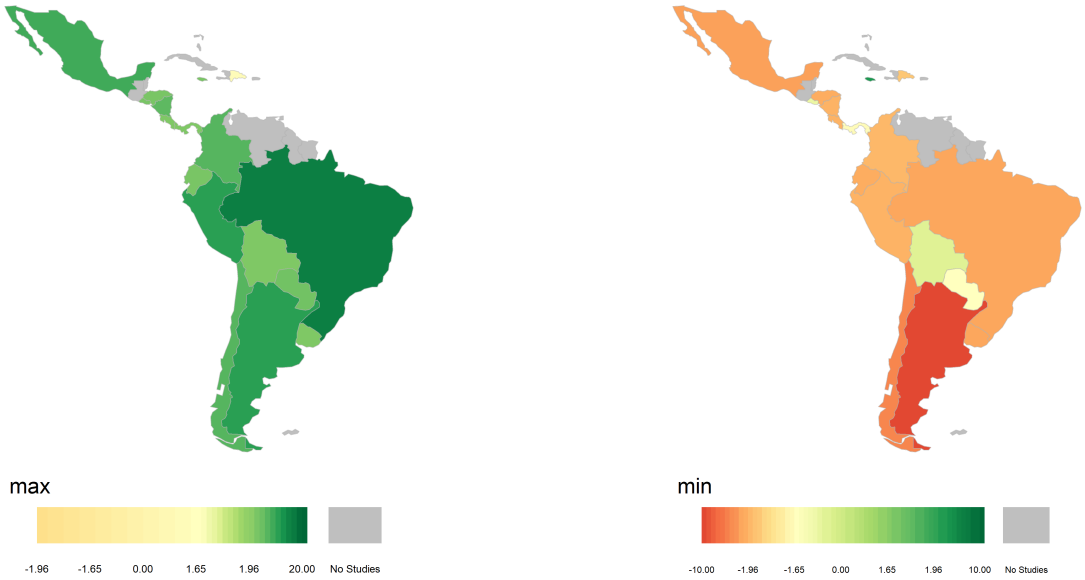
(A) 2000



(B) 2015

FIGURE I.2: Conditional Cash Transfers and evaluation status in low and middle income countries

Notes: A country is classified as having an evaluated CCT if it has an active CCT program that has been evaluated through a program evaluation either before or including 2015. Data sources for countries outside of LAC: Social Assistance in Low and Middle Income Countries database (Barrientos and Villa, 2015), and Bastagli et al. (2016).



(A) Maximum treatment effect across all studies

(B) Minimum treatment effect across all studies

FIGURE I.3: Cash transfers and cumulative studies by 2015 and country

Notes: Distribution of program evaluation outcomes by country, for full set of evaluations available from 2000 to 2015. Panel a): maximum test statistic of headline results for each country. Panel b): minimum test statistic of headline results for each country. The test statistic is defined as the treatment effect divided by the standard error.

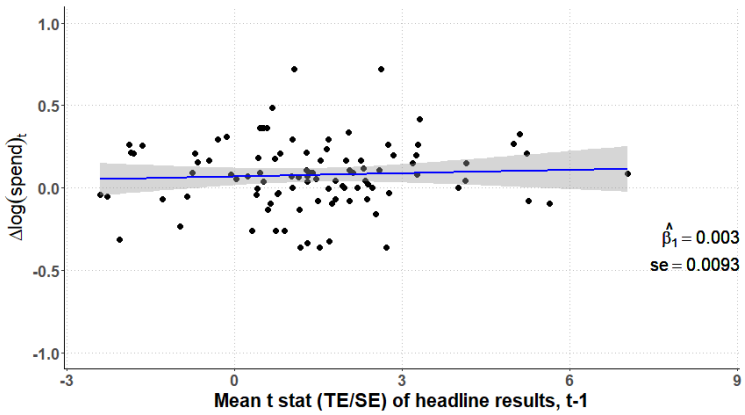


FIGURE I.4: Mean t statistic, and changes in spending

Notes: Linear relationship between causal estimates of impact and changes in spending on the same program, one year after the program evaluation is first available. The evaluation level treatment effect is summarised as the mean of the t-statistic (statistical significance) of headline results.

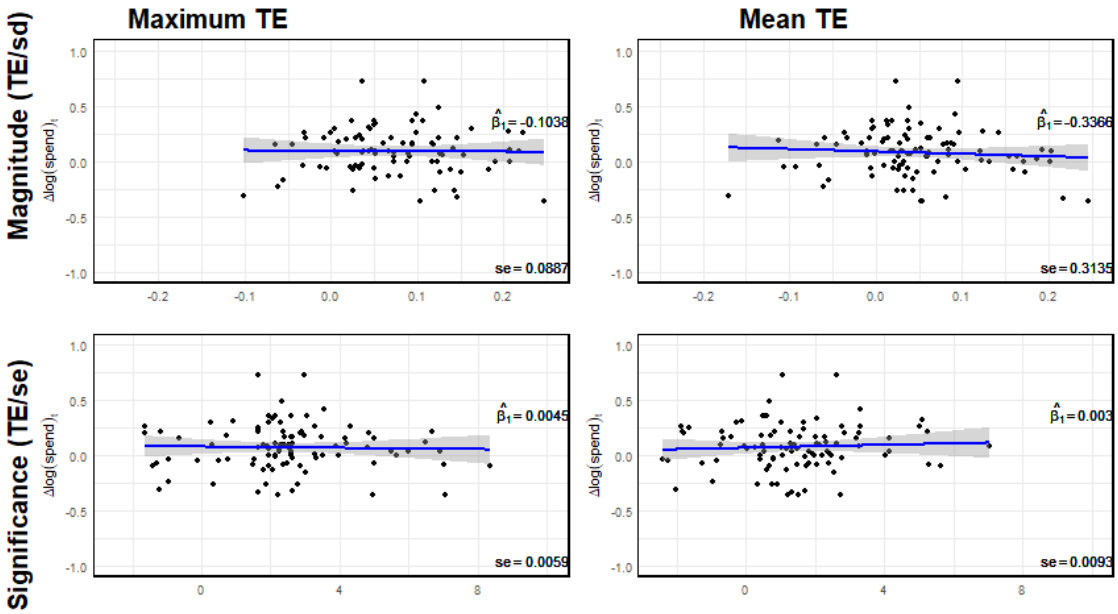


FIGURE I.5: Treatment effects and changes in spending on the same program, across measures of evaluation outcomes

Notes: Linear relationship between causal estimates of cash transfer impact and spending on the same program, across measures of evaluation outcomes. (1) Top left quadrant: maximum magnitude (effect size); (2) Bottom left: maximum significance (t-statistic); (3) Top right: mean magnitude (effect size); (4) Bottom right: mean significance (t-statistic).

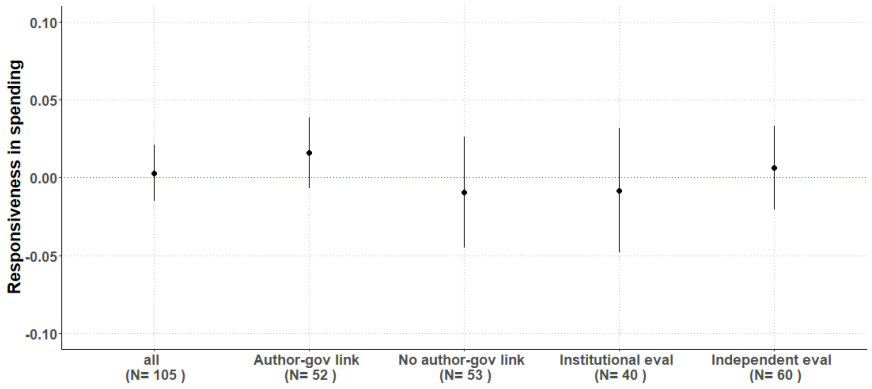


FIGURE I.6: Responsiveness in spending to subsets of evidence, by government-author relationships and source of evaluation

Notes: Linear relationship between evaluation outcomes and CCT spending, and 95% confidence intervals across subsets. *All:* full sample. *Author-gov link:* at least one author has a working relationship with the government; *Institutional evaluation:* demanded by government or international institutions; *Independent evaluation:* demanded and conducted by independent researchers.

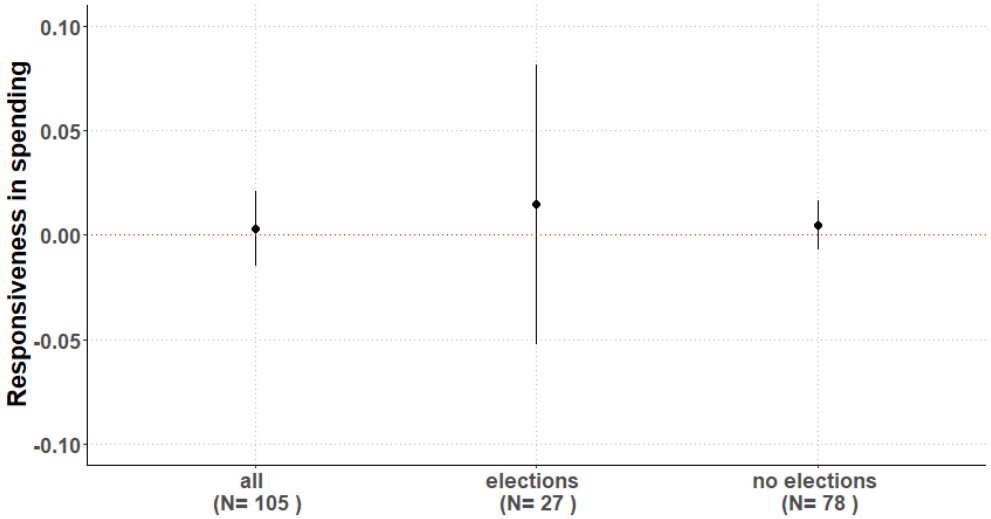
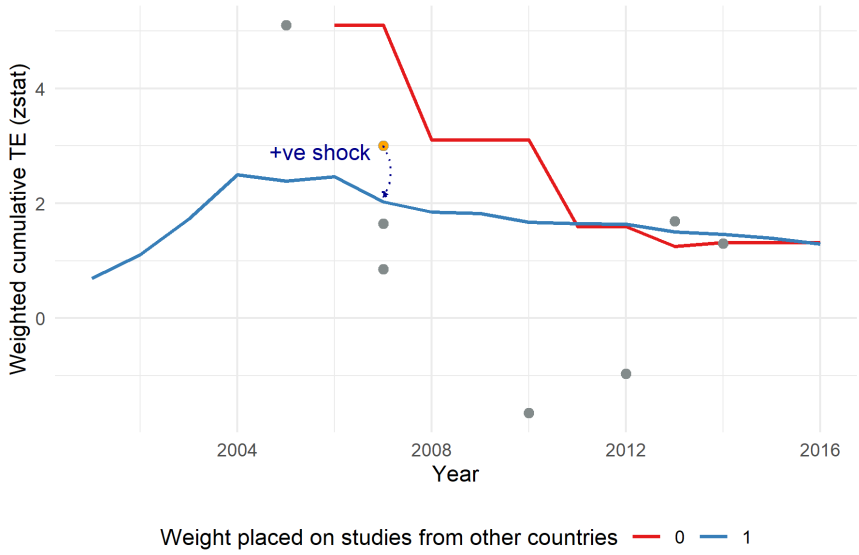


FIGURE I.7: Linear relationship between TE and spending, by political conditions

Notes: Linear relationship between evaluation outcomes and CCT spending, and 95% confidence intervals across subsets. *'All'* refers to the full sample; *'Elections'* refers to evaluations that are first published in an election year; *'No elections'* refers to evaluations that are first published in non-election years.

(A) $\lambda = 0$



(B) $\lambda = 1$

FIGURE I.8: Illustrative example of quantified surprises, by assumptions on external validity

Notes: This figure illustrates how the same evaluation can be interpreted as a positive or a negative surprise, depending on assumptions on λ, the external validity of studies from other countries. Each dot represents a new evaluation. Solid lines represent the estimated cumulative beliefs, based on cumulative evidence across assumptions of zero external validity (Panel a, $\lambda = 0$), and perfect external validity (Panel b, $\lambda = 1$).

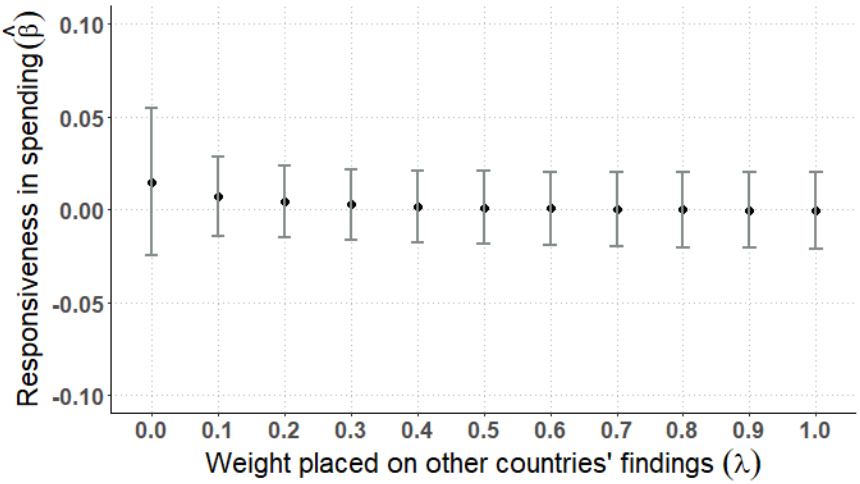
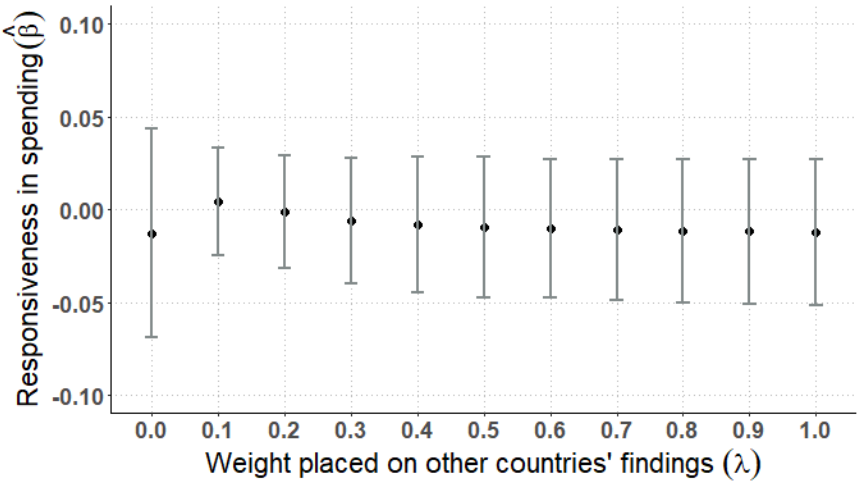
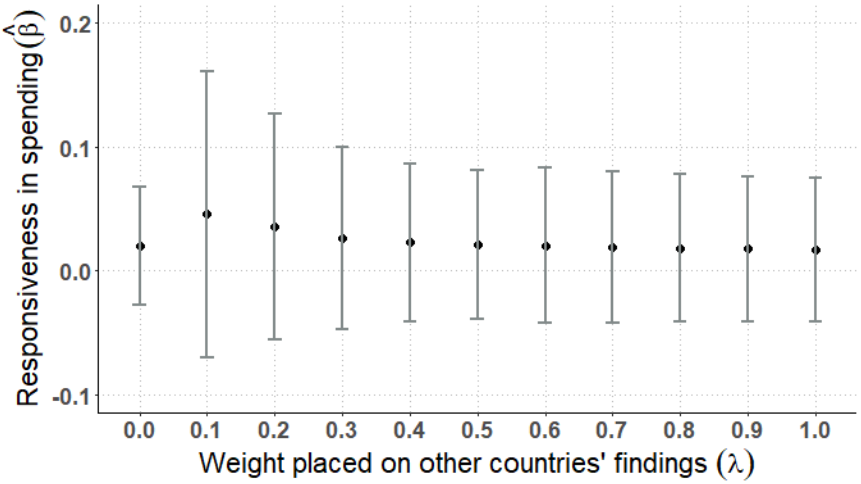


FIGURE I.9: Relationship between quantified surprises and spending, across different assumptions on λ

Notes: Estimated coefficient and 95% confidence intervals, for the linear relationship between quantified surprises and CCT spending, and across assumptions of λ. $\lambda = 0$: beliefs of zero external validity, i.e. zero weight is placed on research results from other countries; $\lambda = 1$ corresponds with beliefs of perfect external validity, i.e. equal weight is placed on research results from all countries.



(A) Negative surprises only



(B) Positive surprises only

FIGURE I.10: Relationship between quantified surprises and spending, across different assumptions on λ. Sample split by negative vs. positive surprises

Notes: This figure plots the estimated coefficient and 95% confidence intervals, for the linear relationship between quantified surprises and CCT spending, and across assumptions of λ. Sample estimated separately for positive surprises and negative surprises.

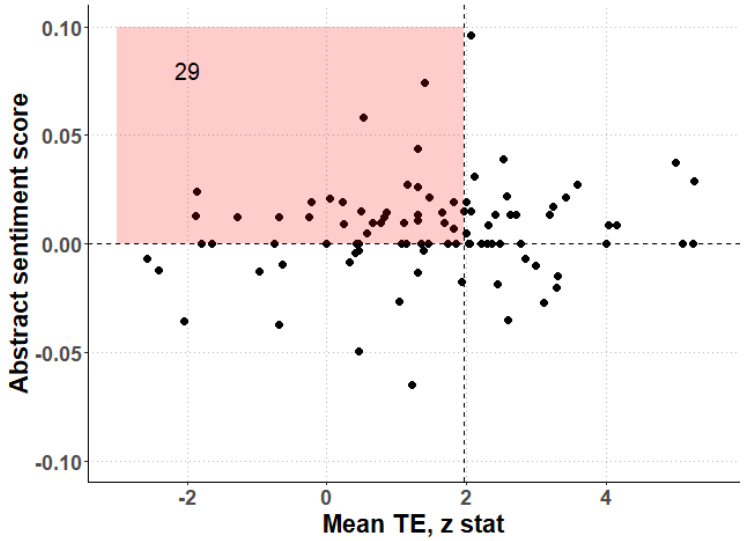


FIGURE I.11: Mean Treatment effect (t-statistic) and the abstract sentiment score

Notes: Abstract sentiment score: difference between the share of positive sentiment words in the abstract and the share of negative sentiment words in the abstract. The red shaded region highlights papers that have a mean null or negative treatment effect (insignificant at the 5% level), and are positively framed in the abstract text.

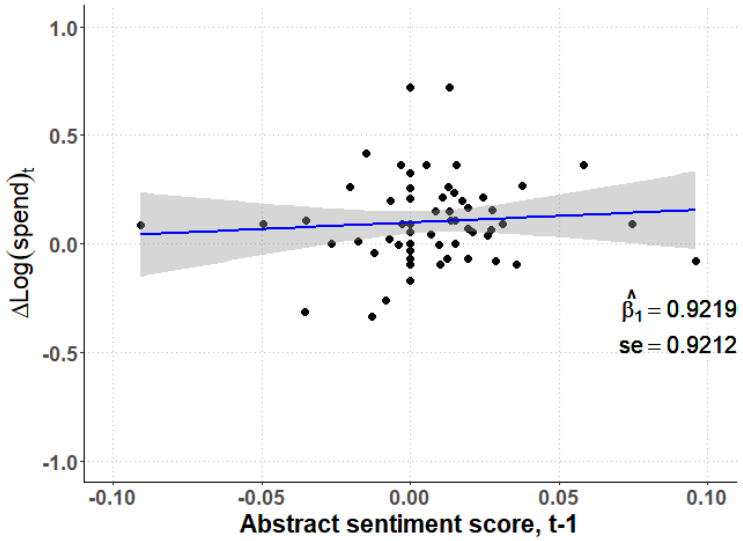


FIGURE I.12: Abstract sentiment score and changes in spending

Notes: Linear relationship between the abstract sentiment score and changes in spending on the same program, one year after the program evaluation is first available. Abstract sentiment score: difference between the share of positive sentiment words in abstract and the share of negative sentiment words in abstract.

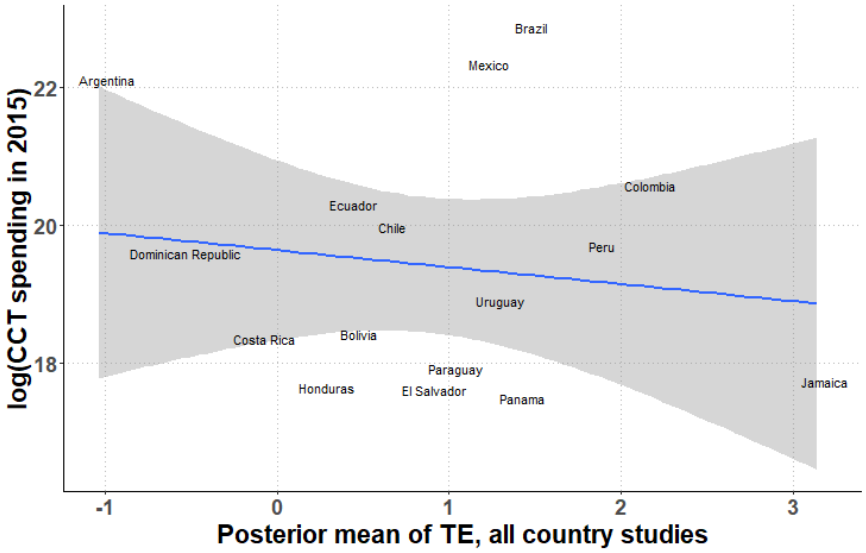


FIGURE I.13: Bayesian posterior mean of aggregate results in 2015, and cash transfer spending

Notes: Posterior mean of the aggregate country level treatment effects, based on all evidence published on CCTs in country *i* between 2000-2015.

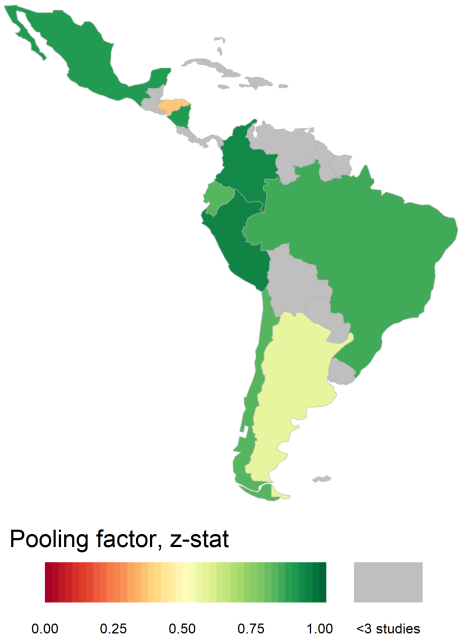


FIGURE I.14: Estimated pooling factor of aggregate studies by country

Notes: Estimated generalized pooling factor for each country, based on all evidence published on CCTs in country *i* between 2000-2015. Excludes countries that have less than three evaluations.

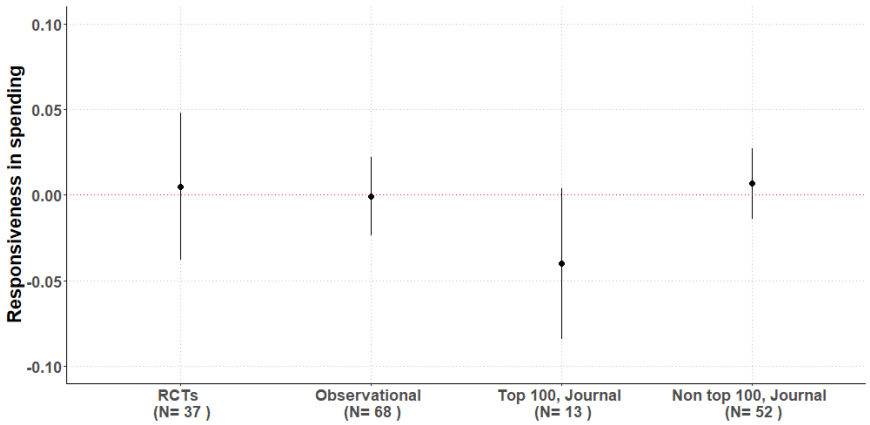


FIGURE I.15: Responsiveness in spending to subsets of evidence, by credibility

Notes: Linear relationship between program evaluation outcomes and changes in spending, across subsets of studies by measures of credibility. *Experimental*: main identification strategy uses experimental variation; *Non experimental*: main identification strategy uses observational methods, e.g. IV, DiD. *Top 100*: evaluation is published in a top 100 academic journal; *Non-top 100* evaluation is not published in top 100 journal.

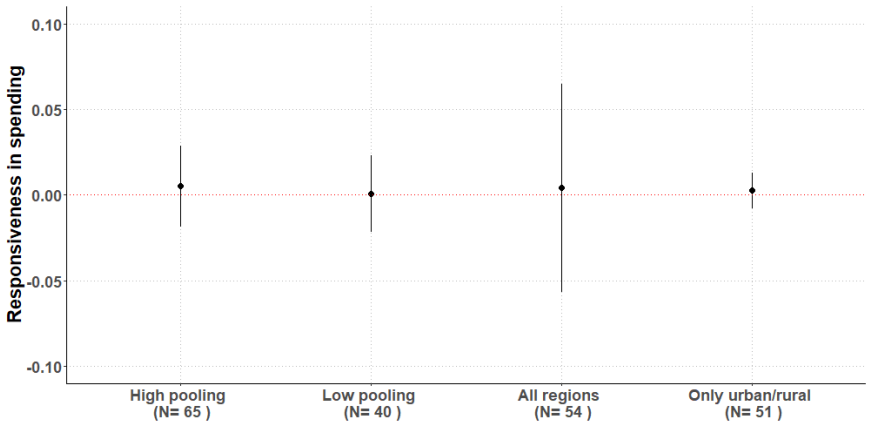


FIGURE I.16: Responsiveness in spending to subsets of evidence, by generalizability

Notes: Linear relationship between program evaluation outcomes and changes in spending, across subsets of studies by measures of generalizability. *High pooling*: estimated pooling factor of the evaluation is higher than 0.6. *Full population*: program evaluations that estimate the treatment effect for the full population, i.e. no sub-region. *Urban/Rural*: program evaluations that estimate the treatment effect only for rural or urban populations.

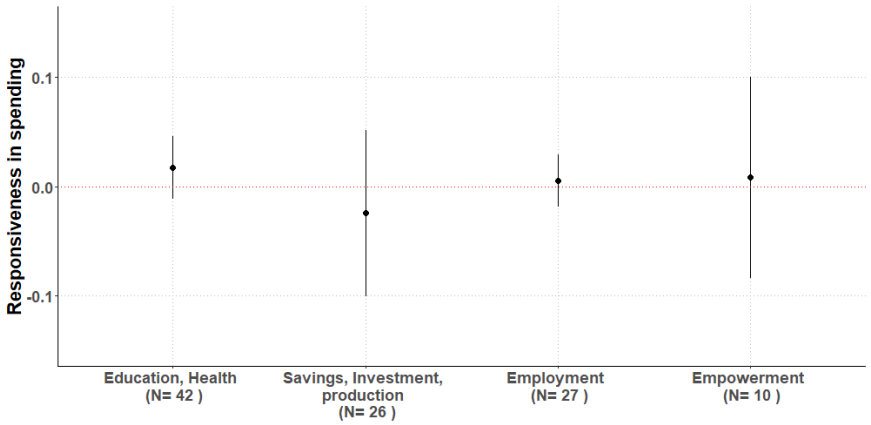


FIGURE I.17: Responsiveness in spending to subsets of evidence, by outcome type

Notes: Linear relationship between program evaluation outcomes and changes in spending, by main outcome of interest in the study.

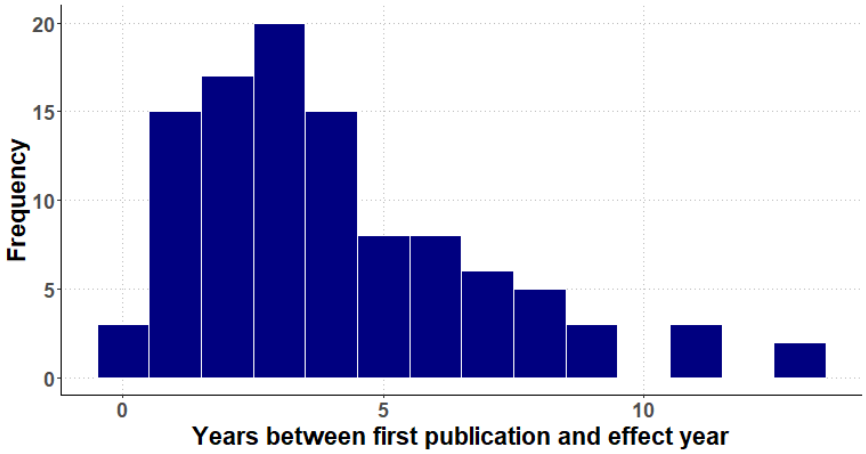


FIGURE I.18: Timeliness of studies: distribution of number of years between the effect year, and the first year of publication

Notes: This figure plots the number of studies by the number of years between the effect year and first year of publication. Effect year: year pertaining to the treatment effect of interest, e.g. the endline year for experimental evaluations, and the post-period for quasi-experimental evaluations

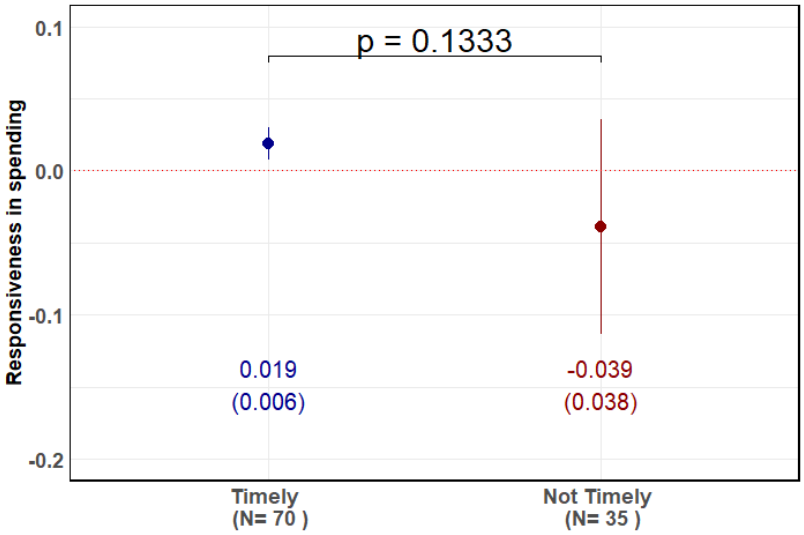


FIGURE I.19: Responsiveness in spending to subsets of evidence, by timeliness

Notes: Linear relationship between evaluation outcomes and changes in spending, by timeliness of evaluation. *Timely*: evaluation is first published within four years of the effect year.

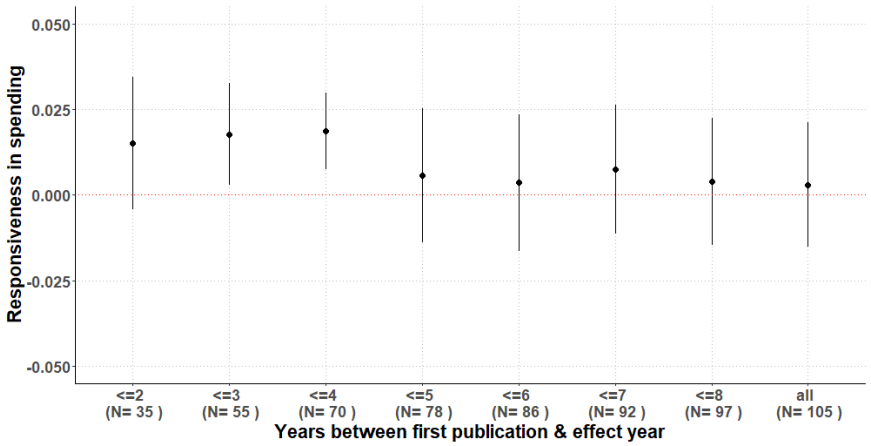


FIGURE I.20: Responsiveness in spending by years between first publication and effect year

Notes: Linear relationship between evaluation outcomes and changes in spending, by number of years between first publication and effect year.

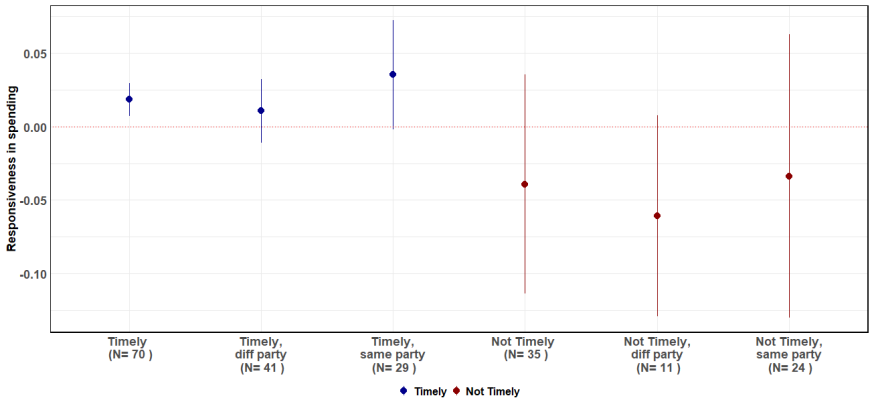
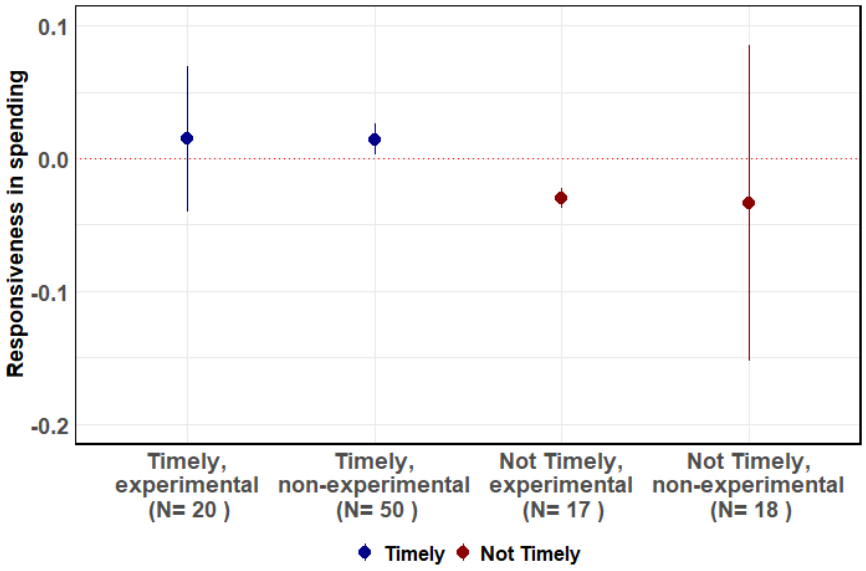
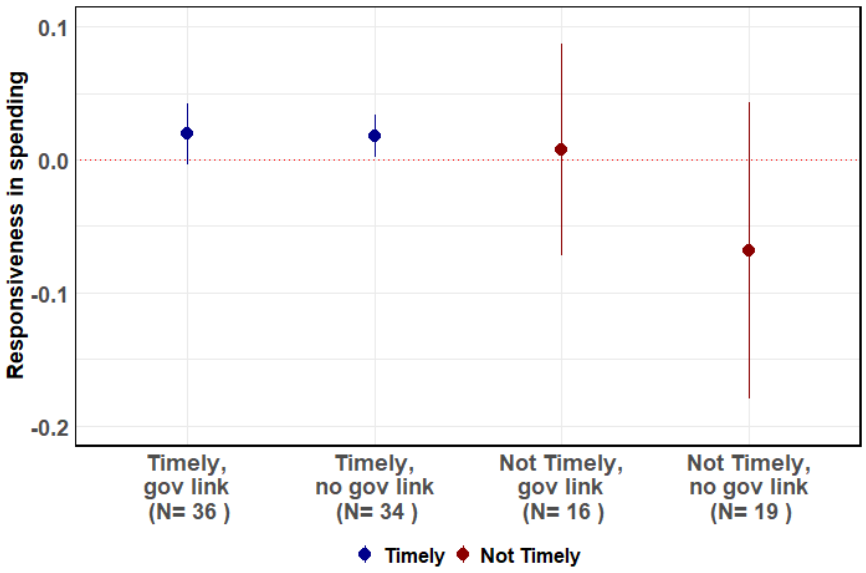


FIGURE I.21: Responsiveness in spending to subsets of evidence, by timeliness and political party in power

Notes: Linear relationship between evaluation outcomes and changes in spending, by timeliness and political party. *Timely*: evaluation is first published within four years of the effect year. *Sameparty*: the political party at the time of the first publication is the same as the party at the time of the effect year. $Diffparty$: the political party at the time of the first publication is the different from the party at the time of the effect year.



(A) Responsiveness in spending by timeliness & methodology



(B) Responsiveness in spending by timeliness & government relationships

FIGURE I.22: Responsiveness in spending to subsets of evidence, by timeliness and other characteristics

Notes: Linear relationship between evaluation outcomes and changes in spending, by timeliness and other characteristics. *Timely*: evaluation is first published within four years of the effect year. *Experimental*: main identification strategy uses experimental variation. *Govlink*: author has a working relationship with the implementing government.

I.9.2 Tables

TABLE I.1: Summary of studies, treatment effects, and methods

	Studies (S)	Treatment effects (N)
Aggregate	128	468
Experimental	50	
Non experimental	79	
Outcome of interest		
Education	53	128
Employment	57	132
Empowerment	13	33
Health & Nutrition	36	79
Monetary poverty	31	57
Savings, Investment, Production	12	39

Notes: This table shows summary characteristics of program evaluations in my sample, by empirical methodology and outcome of interest. The total methods and outcomes of interest do not sum up to the aggregate, because there are multiple impact evaluation that measure multiple outcomes of interest in the same paper; and one paper that uses both experimental and non-experimental variation for different outcome variables of interest.

TABLE I.2: Source of Program Evaluations

	N
Total	128
Author & institutional collaboration	65
Independent evaluation	70
Demanding agent	
Government	30
International institution	25
Independent researcher	70
Evaluating agent	
Government	2
International institution	14
Independent researcher	109

Notes: Author-institutional collaboration: studies where at least one author has a working relationship with the implementing government. *Independent evaluation:* demanding and evaluating agents of the evaluation are both independent researchers. *Demanding agent:* person or organisation who initiated or requested the program evaluation. *Evaluating agent:* person or organisation who conducted the program evaluation.

TABLE I.3: Example of study level summary metrics based on Galiani and McEwan (2013)

	Maximum	**Mean**
Magnitude (effect size)	0.02	0.19
Significance (TE/SE)	3.48	3.1

TABLE I.4: Relationship between mean t-stat and subsequent spending, with country and time fixed effects

	$\Delta log(spend)_{it}$		
	(1)	(2)	(3)
Constant	0.0273		
	(0.0518)		
TE_{it-1}	0.0030	0.0058	0.0040
	(0.0093)	(0.0116)	(0.0124)
country FE		Yes	Yes
time FE			Yes
Fit statistics			
Observations	105	105	105
R^2	0.00027	0.20199	0.35110
Within R^2		0.00108	0.00045

Clustered (country) standard-errors in parentheses

Notes: Linear relationship between causal estimates of impact and changes in spending on the same program, one year after the program evaluation is first available. The evaluation level treatment effect (TE_{it-1}) of a study in country i first made available in year $t-1$, is summarised as the mean of the t-statistic (statistical significance) of headline results. $spend_{it}$ is the aggregate spending on the evaluated cash transfer program in year t .

TABLE I.5: Relationship between measures of evaluation outcomes and spending, one year after first publication of evaluation results

	Dependent variable: $\Delta log(y_{it})$				
	Measure of evaluation outcome				
	Mean t-stat	Max t-stat	Mean effect size	Max effect size	Abstract sentiment
Constant	0.0273	0.0189	0.0481	0.0438	0.0248
	(0.0518)	(0.0653)	(0.0459)	(0.0518)	(0.0821)
TE_{it-1}	0.0030	0.0045	-0.3366	-0.1038	0.9219
	(0.0093)	(0.0059)	(0.3135)	(0.0887)	(0.9212)
Observations	105	105	105	105	64
R^2	0.00027	0.00117	0.01111	0.00597	0.00294
Adjusted R^2	-0.00943	-0.00852	0.00151	-0.00368	-0.01314

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, one year after the program evaluation is first published. The evaluation results (TE_{it-1}) in each study are summarised by: (1) the mean t-statistic of headline results; (2) maximum t-statistic of headline result; (3) mean effect size of headline results; (4) maximum effect size of headline results; and (5) the abstract sentiment. Standard errors are clustered at the country level.

TABLE I.6: Relationship between measures of evaluation outcomes and spending, two years after first publication of evaluation results

	Dependent variable: $\Delta log(y_{i,t+1})$				
	Measure of evaluation outcome				
	Mean t-stat	Max t-stat	Mean effect size	Max effect size	Abstract sentiment
Constant	0.0918	0.0862	0.1182	0.1274	0.0781
	(0.0817)	(0.0906)	(0.0740)	(0.0733)	(0.1078)
TE_{it-1}	0.0211*	0.0116	0.0068	0.0068	0.9877
	(0.0105)	(0.0079)	(0.0084)	(0.0051)	(1.3632)
Observations	101	101	77	77	62
R^2	0.00449	0.00254	0.00549	0.01472	0.00111
Adjusted R^2	-0.00557	-0.00754	-0.00777	0.00158	-0.01554

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, two years after the program evaluation is first published. The evaluation results (TE_{it-1}) in each study are summarised by: (1) the mean t-statistic of headline results; (2) maximum t-statistic of headline result; (3) mean effect size of headline results; (4) maximum effect size of headline results; and (5) the abstract sentiment. Standard errors are clustered at the country level.

TABLE I.7: Relationship between measures of evaluation outcomes and spending, three years after first publication of evaluation results

	Dependent variable: $\Delta log(y_{i,t+2})$				
	Measure of evaluation outcome				
	Mean t-stat	Max t-stat	Mean effect size	Max effect size	Abstract sentiment
Constant	0.2394	0.2152	0.2687	0.2812	0.2464
	(0.0708)	(0.0579)	(0.0757)	(0.0677)	(0.0976)
	(0.0817)	(0.0906)	(0.0740)	(0.0733)	(0.1078)
TE_{it-1}	0.0200	0.0180	-0.0074	-0.0023	4.9961
	(0.0138)	(0.0107)	(0.0088)	(0.0032)	(4.0630)
Observations	98	98	75	75	60
R^2	0.00360	0.00549	0.00610	0.00160	0.02768
Adjusted R^2	-0.00678	-0.00487	-0.00751	-0.01207	0.01091

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, three years after the program evaluation is first published. The evaluation results (TE_{it-1}) in each study are summarised by: (1) the mean t-statistic of headline results; (2) maximum t-statistic of headline result; (3) mean effect size of headline results; (4) maximum effect size of headline results; and (5) the abstract sentiment. Standard errors are clustered at the country level.

TABLE I.8: Relationship between TE_{it-1} and $\Delta log(spend_{it})$, by measures of credibility

	Dependent variable: $\Delta log(spend_{it})$			
	Subset of evaluations			
	Experimental	Non-experimental	Top 100	Non-top 100
Constant	0.0923	-0.0054	-0.0013	0.0183
	(0.0574)	(0.0558)	(0.1336)	(0.0246)
TE_{it-1}	0.0050	-0.0009	-0.0401	0.0067
	(0.0219)	(0.0117)	(0.0224)	(0.0104)
Observations	37	68	13	52
R^2	0.00250	2.06×10^{-5}	0.02984	0.00404
Adjusted R^2	-0.02600	-0.01513	-0.05835	-0.01588

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, one year after the program evaluation is first published, across subsets of credibility. The treatment effect is estimated as the mean of the t-statistic of headline results. *Experimental*: main identification strategy uses experimental variation; *Non experimental*: main identification strategy uses observational methods, e.g. IV, DiD. *Top 100*: evaluation is published in a top 100 academic journal; *Non-top 100* evaluation is not published in top 100 journal. Standard errors are clustered at the country level.

TABLE I.9: Relationship between TE_{it-1} and $\Delta log(spend_{it})$, by measures of generalizability

	Dependent variable: $\Delta log(spend_{it})$			
	Subset of evaluations			
	High pooling	Low pooling	Full population	Urban/Rural
Constant	0.0351	0.0130	0.0305	0.0232
	(0.0560)	(0.1171)	(0.0491)	(0.0622)
TE_{it-1}	0.0052	0.0009	0.0041	0.0026
	(0.0121)	(0.0114)	(0.0310)	(0.0052)
Observations	65	40	54	51
R^2	0.00237	1.37×10^{-5}	0.00038	0.00024
Adjusted R^2	-0.01347	-0.02630	-0.01884	-0.02016

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, one year after the program evaluation is first published, across subsets of generalizability. The treatment effect is estimated as the mean of the t-statistic of headline results. *High pooling*: estimated pooling factor of the evaluation is higher than 0.6. *Full population*: program evaluations that estimate the treatment effect for the full population, i.e. no sub-region. *Urban/Rural*: program evaluations that estimate the treatment effect only for rural or urban populations. Standard errors are clustered at the country level.

TABLE I.10: Relationship between TE_{it-1} and $\Delta log(spend_{it})$, by outcome categories

	Dependent variable: $\Delta log(spend_{it})$			
	Subset of evaluations			
	Education, Health	Savings, Investment, Production	Employment	Empowerment
Constant	0.0192	0.0594	0.0233	0.0628
	(0.0215)	(0.0473)	(0.1104)	(0.1056)
TE_{it-1}	0.0172	-0.0246	0.0055	0.0082
	(0.0146)	(0.0387)	(0.0122)	(0.0472)
Observations	42	26	27	10
R^2	0.01814	0.00899	0.00078	0.00664
Adjusted R^2	-0.00641	-0.03230	-0.03919	-0.11753

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, one year after the program evaluation is first published, across subsets of outcome categories. The treatment effect is estimated as the mean of the t-statistic of headline results. The outcome category of each study is defined using the main outcomes of interest in the headline results. Standard errors are clustered at the country level.

TABLE I.11: Relationship between TE_{it-1} and $\Delta log(spend_{it})$, by timeliness

	Dependent variable: $\Delta log(spend_{it})$	
	Timely evaluations	Not timely evaluations
Constant	0.0547	-0.0039
	(0.0485)	(0.0812)
TE_{it-1}	0.0185***	-0.0392
	(0.0057)	(0.0380)
Observations	70	35
R^2	0.01345	0.02888
Adjusted R^2	-0.00106	-0.00055

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, one year after the program evaluation is first published, by timeliness of the evaluations. The treatment effect is estimated as the mean of the t-statistic of headline results. *Timely*: evaluation is first published within four years of the effect year. Standard errors are clustered at the country level.

	Not timely	Timely
N	35	70
Experimental	0.49	0.29
Top 100 publication	0.20	0.09
Government collaboration	0.46	0.51

TABLE I.12: Characteristics of timely versus not-timely studies

Notes: Timely: evaluation is first published within four years of the effect year. *Experimental*: main identification strategy uses experimental variation. *Govlink*: author has a working relationship with the implementing government.

TABLE I.13: Relationship TE_{it-1} and $\Delta log(spend_{it})$, by timeliness and political party

Dependent variable: $\Delta log(spend_{it})$			
Panel A: Timely Evaluations			
	All	Different party	Same party
Constant	0.0547	0.0117	0.1124*
	(0.0485)	(0.0653)	(0.0607)
TE_{it-1}	0.0185***	0.0106	0.0353*
	(0.0057)	(0.0110)	(0.0189)
Observations	70	41	29
R^2	0.01345	0.00334	0.12506
Adjusted R^2	-0.00106	-0.02222	0.09265
Panel B: Not Timely Evaluations			
	All	Different party	Same party
Constant	-0.0039	0.0998	-0.0448
	(0.0812)	(0.1857)	(0.0457)
TE_{it-1}	-0.0392	-0.0608	-0.0336
	(0.0380)	(0.0349)	(0.0492)
Observations	35	11	24
R^2	0.02888	0.09796	0.01996
Adjusted R^2	-0.00055	-0.00227	-0.02459

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, one year after the program evaluation is first published, by timeliness of the evaluations. The treatment effect is estimated as the mean of the t-statistic of headline results. *Timely*: evaluation is first published within four years of the effect year. *Sameparty*: the political party at the time of the first publication is the same as the party at the time of the effect year. Standard errors are clustered at the country level.

TABLE I.14: Relationship between TE_{it-1} and $\Delta log(spend_{it})$, by timeliness and other characteristics

Dependent variable: $\Delta log(spend_{it})$					
Panel A: Timely Evaluations					
	All	Experimental	Non-experimental	Govlink	No govlink
Constant	0.0547	0.1343	0.0291	0.0532	0.0564
	(0.0485)	(0.1007)	(0.0481)	(0.0866)	(0.0429)
TE_{it-1}	0.0185***	0.0148	0.0144**	0.0192	0.0177*
	(0.0057)	(0.0280)	(0.0062)	(0.0116)	(0.0082)
Observations	70	20	50	36	34
R^2	0.01345	0.02541	0.00659	0.00912	0.03753
Adjusted R^2	-0.00106	-0.02873	-0.01411	-0.02003	0.00745

Panel B: Not Timely Evaluations					
	All	Experimental	Non-experimental	Govlink	No govlink
Constant	-0.0039	0.0612	-0.0845	-0.0461	0.0237
	(0.0812)	(0.0368)	(0.1069)	(0.0637)	(0.1301)
TE_{it-1}	-0.0392	-0.0299**	-0.0335	0.0072	-0.0681
	(0.0380)	(0.0038)	(0.0607)	(0.0406)	(0.0566)
Observations	35	17	18	16	19
R^2	0.02888	0.08983	0.01306	0.00311	0.06061
Adjusted R^2	-0.00055	0.02916	-0.04863	-0.06810	0.00535

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and changes in spending on the same program, one year after the program evaluation is first published, by timeliness of the evaluations. The treatment effect is estimated as the mean of the t-statistic of headline results. *Timely*: evaluation is first published within four years of the effect year. *Experimental*: main identification strategy uses experimental variation. *Govlink*: author has a working relationship with the implementing government. Standard errors are clustered at the country level.

Chapter II

Gender differences in altruism: a Bayesian hierarchical analysis of dictator games

Abstract: I aggregate evidence on gender differences in dictator game giving from experiments published in all working papers and peer-reviewed journals since 1990. Using a two-stage Bayesian hierarchical model, I find that on average women give around 3 percentage points more than men in studies of dictator games. I show that while this estimate is smaller than that found in previous studies, it is likely to be an upper bound estimate due to publication bias. Using a truncated selectivity model, I estimate the conditional probability of publication as a function of experiment results. My findings suggest that experiments that find positive results (i.e. women contribute more than men) and are statistically different from zero at the 5% level, are around 13 times more likely to be published than statistically significant and negative results.

II.1 Introduction

Are women more altruistic than men? Given that assumptions on preferences are central to models of individual choice, gender differences in altruism would have far-reaching implications for theoretical and empirical work in economics. For instance, differences in altruism could explain differences in labour market outcomes between men and women, including in wages and occupational choice (Bertrand, 2011b; Buser et al., 2014). Recent evidence suggests that these differences also matter at the aggregate level, with gender differences in altruism predicting economic development and gender equality across countries (Falk et al., 2016; Falk and Hermle, 2018). With this motivation in mind, a large

body of evidence measures altruism using lab and field experiments. Yet the overall findings from this literature are ambiguous and inconclusive.

In this paper, I study whether women are more altruistic than men by aggregating evidence from first-mover behaviour in dictator games. I collect data on gender differences in dictator game behaviour from all working papers and journals, regardless of whether or not gender was the main topic of interest. My sample covers results on gender differences in giving from 100 dictator games across 35 studies and represents the decisions of 20,265 participants. Considered individually, the conclusions from these experiments are mixed. While Eckel and Grossman (1998), Andreoni and Vesterlund (2001), and Boschini et al. (2018) find that women give more than men when the price of giving is one, Bolton and Katok (1995), Ben-Ner et al. (2004) and Cadsby et al. (2010) find limited evidence for gender differences. Extrapolating a general finding from these studies is difficult since differences in results are likely driven by both sampling variation and genuine variation in experimental design and characteristics.

Using a Bayesian hierarchical model, I quantify the overall giving of women relative to men in dictator games. Compared with classical approaches to meta-analysis, the Bayesian hierarchical model allows me to jointly estimate the overall gender differences in dictator game giving, and the heterogeneity across studies. This allows me to separate between within-study and across-study variation, and consequently, to estimate the extent to which findings from one study can help us learn about the overall population effect. My approach complements the growing literature in economics that uses Bayesian hierarchical models to aggregate findings across contexts (e.g. Burke et al., 2015; Bandiera et al., 2016; Vivalt, 2016; Meager, 2019).

My findings suggest that women give 3.2 percentage points more of their endowment than men, with 95 percent probability that the true mean is between 1.7 and 4.5 percentage points. Using pooling metrics suggested by Gelman et al. (2006), I find that on average 89 percent of the heterogeneity in effects across studies is explained by sampling variation. Thus, genuine heterogeneity across studies is low and each additional study is likely to be informative of the overall population effect.

I then turn my attention to exploring publication bias. The results from the Bayesian hierarchical model can be interpreted as a best estimate of gender differences in giving, within the context of dictator game experiments that report results on gender and are available in working papers and journals. The extent to which these findings generalise to a broader sample relies on how representative published papers are of dictator game giving in the overall population. If, for instance, papers that find that women give more than men are more likely to be published, then the estimated result from the Bayesian hierarchical model is likely to be an overestimate of the general population effect.

Using a truncated selectivity model, I parametrically estimate the conditional probability of publication, where following Andrews and Kasy (2019), I assume that the publication decisions of researchers and editors are a function of the study results. My results are strongly suggestive of selective publication. Overall, papers that find a significant and positive (i.e. women give more than men) result are over 13 times more likely to be published than papers that find a statistically significant and negative result.

I find evidence that the selection rule is complex, and differs by the topic of the paper and the quality of the journal. Among papers that explicitly study gender, I find evidence for selection based on statistical significance, but not on the sign of the effect. Among high-quality peer-reviewed journals,[1] I find that positive and significant results are more likely to be published than positive and insignificant results.

Taken together, the findings from the Bayesian hierarchical model suggest that women give more than men in dictator games in the context of studies where gender results are made available. While this result is smaller than that found in existing studies, it is likely to be an upper bound of the overall population effect since results that are positive and significant are more likely to be published.

My findings relate to the literature on gender differences in social preferences by aggregating findings from dictator games. Existing review articles provide a qualitative assessment of the literature (e.g. Eckel and Grossman, 2008; Croson and Gneezy,

[1] I measure high-quality as journals with 5-year average impact factors ranked in the top two quartiles in the Annual Journal Citation Reports.

2009a). I contribute to these findings by estimating the average differences in giving and quantifying the likely heterogeneity across studies. Relatedly, Engel (2011) uses classical meta-analysis techniques to analyse findings from dictator game experiments. However whereas Engel (2011) separately estimates the average effect and the cross-study heterogeneity and therefore likely underestimates heterogeneity (Rubin, 1981), I am able to jointly estimate these two variables of interest using Bayesian hierarchical methods.

Finally, I highlight a new reason for the gender gap in giving observed in existing papers: publication bias. My findings here are related to the growing economics literature on publication bias (Simonsohn et al., 2014; Brodeur et al., 2016; Andrews and Kasy, 2019). Various other reasons for gender differences in dictator games have been suggested in the literature, including the price of giving (Andreoni and Vesterlund, 2001), gender priming (Boschini et al., 2018), and anonymity of decision-making (Dufwenberg and Muren, 2006). I stress that while these experimental differences are potential sources of gender differences in giving, in the presence of selective publication, findings from the literature are likely to overestimate the differences in the overall population.

II.2 Data and Context

II.2.1 Selection of studies

To study gender differences in altruism, I focus my attention on behaviour in dictator games. Introduced by Forsythe et al. (1994) and Kahneman et al. (1986), the dictator game is a lab experiment involving two players, often referred to as the proposer and the recipient. The proposer is given a sum of money and decides what proportion of the money to offer to the recipient, versus what proportion to keep for themselves. For rational and purely self-interested agents, the subgame perfect Nash equilibrium of the dictator game is for proposers to keep the entire sum of money for themselves (i.e. to offer zero). Thus, a positive offer in the first stage is often interpreted as evidence for altruistic preferences.

While altruism is sometimes measured using other lab experiments, such as public goods games or ultimatum games, the dictator game is arguably the cleanest experimental measure for altruism because it is simple and involves limited strategic interactions (Camerer and Fehr, 2004; Eckel and Grossman, 2008). Relatedly, first-mover behaviour in the dictator game remains one of the most prominent measures for altruism in the lab, and has been shown to predict individual and aggregate economic outcomes (Becker et al., 2011; Falk et al., 2016; Falk and Hermle, 2018).

I collect data on all relevant dictator games published in working papers and journals published up until the end of 2019 (when this data was collected). I use two data sources and approaches for compiling relevant papers depending on whether the paper was published prior to or post 2010.

For papers published from 2010 onwards, I conduct a keyword search of "dictator game" and the phrases, "altruism", "generosity", "philanthropy", or "intergenerational transfers" on two databases, EconLit and RePEc. This leaves me with a total of 328 unique papers from journals and working papers published from 2010 to 2019.

To narrow down the search to a relevant sample, I focus my study on non-interactive, one-stage dictator games. Common variants of dictator games include giving recipient power, adding multiple stages, or requiring effort to generate the endowment. As such variants are often designed to measure preferences other than altruism (e.g. risk preferences, reciprocity) I do not include them in my sample.

Of the relevant experimental designs, I include papers in my study if the authors report the average giving in a dictator game of men and women (or the difference between the two) and their associated standard errors; or if the full data is provided, such that these values can be calculated. Note that while most experiments collect data on gender, contributions disaggregated by gender are often not reported. In fact, 6 studies in my search state that there are no gender differences in dictator game giving observed in their experiments, but do not state the average differences in giving, or the standard errors of these differences. Since these papers have to be excluded from my sample, the results

from this study are likely to be an overestimate of the true treatment effect. The selection criteria are summarised in Table II.1.

Of the 328 papers identified in my search, 23 of the papers report average giving in a dictator game of men versus women (with corresponding standard errors); and 4 provide raw data on their experiments, which allows me to calculate the required results. This leaves me with 27 relevant papers for papers published from 2010.

For papers published prior to 2010, I use data from Engel (2011)'s meta-analysis of dictator games, which includes all working and published papers on dictator games available on EconLit and RePEC up to the end of 2009. Of the 131 papers in his full dataset, 14 studies provide data on average giving of women versus men in dictator games, and their associated standard errors. I review the studies using the selection criteria defined in Table II.1, and validate the data against the final reported treatment effects of the papers. Of the 14 papers included in Engel (2011), I exclude the results from 6 papers: 2 papers that have since released new versions or have been published in journals after 2009 (and hence are already in my sample); and 4 papers, since they do not meet my selection criteria.

II.2.2 Summary statistics

My final sample comprises results from 35 studies over a total of 100 experiments. A summary of these results is provided in Table II.2. On aggregate, the experiments in my sample cover 20,265 distinct allocation decisions in dictator games, of which 53% are decisions by women. On average women contribute 2.7 percentage points more than men, with men contributing 29.7% and women contributing 32.4% of their endowment. The average giving in my sample irrespective of gender is 31.1%, which is broadly consistent with the average giving found in the literature (for instance, Engel (2011) finds an average giving of 28.35%).

As is common in lab experiments, most studies in my sample have multiple variants of the dictator game within the same paper in an attempt to disentangle how different

experimental characteristics may affect average giving (see: Table B2). Common variants of the dictator game include variation in the price of giving (e.g. Andreoni and Vesterlund, 2001; Visser and Roelofs, 2011); the anonymity of the dictator and identity of the recipient (e.g. Cadsby et al., 2010; Dufwenberg and Muren, 2006; Slonim and Garbarino, 2008); gender priming (e.g. Boschini et al., 2012, 2018); and the framing of the game (e.g. Smith, 2015; van Rijn et al., 2019). There is also variation in the characteristics across studies including differences in location and subject population. 20 out of 38 studies in my sample explicitly mention gender (or a related term) as the main topic of their paper. The majority of the experiments are conducted among a population of university students (27 out of 38 studies).

To control for quality, I use a subset of this full sample that are published in *RelevantJournals*[2] for my baseline analysis. In particular, I include papers published in the top 5 economics peer-reviewed journals and the main field journals in behavioural and experimental economics. I also include papers published in the NBER Working Papers series, the IZA Discussion Papers, and the CEPR Discussion Paper series. This leaves me with results from 83 experiments across 29 studies.

In Figure II.1, I plot the average contribution of men versus women in dictator games, disaggregated by journal type. Results are closely distributed around the 45 degree line, with marginally more study estimates finding higher contributions by women as compared to men. Results from experiments that are not from my *RelevantJournals* list tend to be noisier and at more extreme values than that found for results published in journals on my list. In Section II.5, I explore these relationships more systematically, by estimating how the type of results published may differ by the characteristics of the study.

[2]Full list of Relevant Journals: American Economic Review, Econometrica, Journal of Political Economy, The Quarterly Journal of Economics, Review of Economic studies, Journal of Behavioral and Experimental Economics, Experimental Economics, Journal of Economic Behavior and Organisation, Games and Economic Behavior, Economic Journal, American Economic Journal: Applied Economics, Journal of Economic Psychology, Management Science, NBER Working Paper series, IZA Discussion Papers, CEPR Discussion Paper series

II.3 Methodology

II.3.1 Bayesian hierarchical models

In understanding gender differences in giving in the overall population, the main empirical challenge is in how we should be summarising evidence across different experiments and studies. The estimated difference in giving of women versus men ranges from -0.292 to 0.465. As illustrated in Section II.2.2 however, this range in estimates could be driven by genuine variation across studies and experiments, due to differences in experimental design and setting. Or alternatively, these differences could be driven by sampling variation in the estimate, specific to either the study or the experiment. Bayesian hierarchical models provide a method to disentangle between these two sources of variation, and have been increasingly used in economics (e.g. Rubin, 1981; Bandiera et al., 2016; Vivalt, 2016; Meager, 2019). By separating between these sources of variation, the methodology allows us to obtain improved estimates of the treatment effect within each study, as well as an estimate of the overall treatment effect for the population.

Let $\hat{y}_{jk}$ denote the estimated difference in giving of women relative to men in experiment k within study j, such that a positive effect estimate $\hat{y}_{jk}$ is an instance in which the average giving of women is higher than that of men. In my sample, I have a set of $\hat{y}_{jk}$ and their associated standard errors, $\hat{se}_{jk}$, across $j = 1, 2, ..., J$ studies, where each study has $k = 1, 2, ..., K$ experiments.

To set ideas, consider the following hierarchical model for the data:

$$\begin{aligned} \hat{y}_{jk} &\sim N(y_{jk}, \hat{se}^2{}_{jk}) \quad j = 1..., J, k = 1, ..., K \\ y_{jk} &\sim N(\mu, \tau^2) \end{aligned} \tag{II.1}$$

In this model, each experiment k within study j, obtains an estimate of the average treatment effect, $\hat{y}_{jk}$. This estimated treatment effect is normally distributed around the true mean effect of the experiment, y_{jk}, and has known variance, $\hat{se}^2_{jk}$. In turn, each ex-

periment mean, y_{jk}, is drawn from a common distribution that is normally distributed around the population mean, μ, and variance, τ^2.

This simple model provides a structure to understand differences between the overall effect for the population, μ, and the estimated effects for a given experiment, $\hat{y}_{jk}$. It distinguishes between statistical sampling variation, as captured by $\hat{y}_{jk} - y_{jk}$, and genuine variation in the treatment effect, $y_{jk} - \mu$. The model further nests approaches used in classical meta-analysis, as well as contrasting views on how we should be aggregating results across experiments. On the one hand, when $\tau = 0$, the hierarchical model corresponds to the 'full pooling', classical fixed effects model, where we assume that each experiment is estimating a common population effect. In this case, the best estimate of the overall population mean is a weighted average of the estimated treatment effect per experiment, where each estimate is weighted by its precision ($1/\hat{se}_{jk}^2$). At the other extreme, when $\tau = \infty$ the model corresponds to the 'no pooling' case and returns the original experiment estimates [3]. Under no pooling, we assume that each experiment is estimating its own context specific effect, and hence there is no learning to be done across studies. The hierarchical model is a compromise between these two extremes. The estimated τ^2 gives us a measure of the external validity of study results: intuitively, the smaller is τ, the more each additional experiment estimate, $\hat{y}_{jk}$, tells us about the overall population effect, and hence, the more we should be updating our estimate and beliefs of μ.

The Bayesian hierarchical model builds on the hierarchical model by treating μ and τ as random variables, and assigning distributional assumptions on these variables. This gives us several advantages beyond the hierarchical approach. In treating μ and τ as random variables, we are less likely to underestimate cross-study heterogeneity (Rubin, 1981). The Bayesian approach further allows us to obtain posterior distributions, so that we can obtain probability distributions on our parameter estimates.

The Bayesian hierarchical model is based on four key assumptions: (1) Normality of the estimated experiment effects, $\hat{y}_{jk}$, given parameters y_{jk} and $\hat{se}_{jk}$, where the variance is assumed to be known; (2) Normality of the study-specific mean, y_{jk}, given μ and τ; (3)

[3] In practice, when τ is very large (i.e. more than 5 times the standard error) there is also no pooling.

Exchangeability of the joint distribution of $\{y_{jk}\}_{j=1,k=1}^{J,K}$; and (4) Distributional assumptions on the hyperpriors, μ and τ. I elaborate upon and discuss each of these assumptions in turn.

Assumption (1), the normality of $\hat{y}_j$, follows almost directly from the assumption of internal validity of the inference within each study. Given the sample sizes are all sufficiently large, the central limit theorem justifies the normal distribution. Justification for Assumption (2), is less straight forward, but there are a number of reasons for which normality of the experiment level means is a natural assumption for this analysis. From a frequentist perspective, Efron and Morris (1977) show that under the assumption of normality, shrinkage estimators have smaller mean squared errors than estimators with full pooling. More broadly, McCulloch and Neuhaus (2011) show that inference on μ and τ under the assumption of normality is still generally reliable even when the true underlying distribution is non-normal. From a practical standpoint, the normal-normal hierarchical structure facilitates comparability of estimates with results from classical meta-analyses Gelman et al. (2013), which enables me to compare findings from my analysis to that of Engel (2011).

The third assumption required for the model is that of exchangeability. The data is exchangeable if the joint distribution of $\{y_{jk}\}_{j=1,k=1}^{J,K}$ is invariant to different permutations of the indices. That is, prior to seeing the effect estimates, there is no prior reason to believe that the average contribution of women relative to men would be larger, smaller, or of similar magnitude in any experiment or study versus that of another. In the absence of information to distinguish between the data and effect estimates, Gelman et al. (2013) argue that exchangeability is the best assumption for modelling. When data is available to distinguish between observations, we can structure the model and condition on groups and study characteristics, so that the model instead relies on conditional exchangeability, rather than full exchangeability.

In the context of this paper, there are several potential threats to exchangeability. First and foremost, within each study j, there are k experiments that each provides a distinct estimate of the treatment effect, $\hat{y}_{jk}$. Since experiments within the same study are con-

ducted and designed by the same set of researchers, the effect estimates are likely to be subject to experimenter effects (Rosenthal, 1976). As such, the prior distribution of experiment effects, y_{jk}, within the same study j, are unlikely to be exchangeable.

To account for this, I adopt a two-stage estimation process outlined in Model II.2. In the first stage of the analysis, I obtain study-level effect estimates, $\hat{y}_j$, for all studies with more than one experiment[4]. The first stage Bayesian hierarchical model gives me an effect estimate and associated standard error for each study j. In the second stage of the analysis, I run the full Bayesian hierarchical model on the study-level estimates, $\hat{y}_j$, using either (1) the estimated treatment effect and associated standard error from the first stage if a given study j has more than one relevant experiment, $k > 1$; or (2) the estimated treatment effect and associated standard error reported in a study if the original study has just one relevant experiment ($k = 1$).

First stage: Obtain estimates for $\hat{y}_j$ for $j = 1...J$. For studies where $k = 1$, $\hat{y}_j = \hat{y}_{jk}$. For studies with $k > 1$, $\hat{y}_j$ is the Bayes estimator from:

$$\begin{aligned} \hat{y}_{jk} &\sim N(y_{jk}, \hat{se^2}_{jk}) \quad k = 1...K \\ y_{jk} &\sim N(y_j, se^2{}_j) \end{aligned} \tag{II.2}$$

Second stage: Using posterior means of $\hat{y}_j$ and $\hat{se}_j^2$ from the first stage, estimate:

$$\begin{aligned} \hat{y}_j &\sim N(y_j, \hat{se^2}_j) \quad j = 1...J \\ y_j &\sim N(\mu, \tau^2) \end{aligned}$$

Within the two-stage framework, the assumption of exchangeability now applies to the joint distribution of $\{y_{jk}\}_{k=1}^{K}$ within the same j (in the first stage), and the joint distribution of $\{y_j\}_{j=1}^{J}$ (in the second stage). In a given study, however, there are often variants of lab experiments designed explicitly to tease out gender differences in giving.

[4]Note here that if $k = 1$ for a given j, $\hat{y}_j = \hat{y}_{jk}$.

For instance, in one of their experiments Boschini et al. (2018) remind respondents of their gender prior to playing the dictator game, citing findings from economics and psychology that find evidence of women being more responsive to gender priming and gender stereotypes (e.g. Steele and Aronson, 1995; Benjamin et al., 2010). In these instances, exchangeability of effect estimates is likely to be violated since prior to seeing the data, we would expect the relative giving of women to be higher in experiments with priming, than that of experiments without. Consistent with this reasoning, I exclude all experiments that use priming (e.g. gender, guilt), and 'take' framing (as opposed to 'give')[5] from my main baseline sample.

Finally, to close the model, I specify a prior distribution for the hyperparameters (Assumption (4)). In context of the two-stage estimation, this means that I need to specify prior distributions for $(\hat{y}_j, \hat{se}_j^2)$ in the first stage, and (μ, τ) in the second stage. I use the following prior distributions:

$$\begin{aligned} y_j &\sim N(0, 0.2) \\ se_j &\sim N(0, 0.2) \\ \mu &\sim N(0, 1) \\ \tau &\sim N(0, 1) \end{aligned} \tag{II.3}$$

Where possible I use weakly informative priors, so that the information in the likelihood dominates and the prior distribution has minimal influence on the posterior distribution. However, as noted by Gelman et al. (2017), the prior distribution will matter for posterior inference when the data is weak. This is particularly relevant in the first stage, where we only have a limited set of experiments per study. I thus adopt a 'tighter' distributional assumption in the first stage; whereas, in the second stage, I can use a relatively weaker prior, in line with the fact that I have stronger data.

[5]The 'Take' frame asks the dictator how much money they want to 'take' from the recipient, as opposed to the standard dictator game, which asks how much they want to 'Give'.

II.4 Baseline results

My baseline sample includes all experiments published in relevant journals, other than those that use priming (e.g. gender, guilt), and 'take' framing (as opposed to 'give'). This leaves me with 69 experiments across 29 studies. I summarise this data in Table II.4. The mean gender difference in giving is smaller than that in the full sample (1.3 vs 2.7 percentage points), which follows mechanically from the fact that I have excluded experiments that are designed explicitly to accentuate these gender differences.

I estimate the two-stage model using my baseline data. Figure II.2 summarizes the posterior distribution of the estimated overall effect, μ. On average, women give 3.2 percentage points more than men in dictator games, with 95% probability that the true mean lies between 1.7 and 4.5 percentage points. In section B.1 of the Appendix, I show that these results are robust to different assumptions.

To investigate whether my results are driven by sample selection, in Table II.4, I estimate the model with two other subsets of my sample: (1) the full dataset, with results from all experiments and studies, irrespective of experimental design or journal of publication, and (2) the 'Vanilla' subset, for which I include only standard, one-shot dictator games (i.e. where dictators and recipients are anonymous and the price of giving is equal to one) that are published in my list of relevant journals.

The estimates of the posterior effect remain reasonably stable across all three subsets of the data, and critically, the 95% intervals are positive and do not include zero for any of the subsets. Compared with previous meta-analyses, the estimated difference in contributions is noticeably smaller. For instance, using the random-effects model in a meta-analysis of dictator games, Engel (2011) finds that women give 5.8 percentage points more than men. However, compared to the Bayesian hierarchical model, the random-effects model treats priors, μ and τ, as fixed once estimated and hence likely underestimates cross-study heterogeneity, and overestimates the population effect (Rubin, 1981). Indeed, the estimate of 5.8 percentage points is not included in the 95% probability inter-

val for my Baseline and Full sample datasets [6].

As alluded to in Section II.3.1, the Bayesian hierarchical model gives us some indication of external validity by separating between sampling variation within studies and genuine variation across studies. Intuitively, if the genuine heterogeneity across studies is small or at the extreme, if $\tau = 0$ (corresponding to full pooling), then each study is implicitly estimating a common population effect, μ. In this case, pooling together data across studies not only improves our understanding of the common population effect, μ, but also improves our estimate for the study-specific effect, $\hat{y}_j$. In contrast, if heterogeneity across studies is large, or at the extreme if $\tau = \infty$ (corresponding to no pooling), then each study is estimating a separate independent phenomenon and should be considered in isolation. The degree of genuine variation across studies thus provides an indication of the degree to which we can generalise and learn across contexts.

As the scale of τ is difficult to interpret and compare across contexts, a common measure of cross-study heterogeneity is instead the pooling metric suggested in Gelman and Pardoe (2006), which measures the genuine variation across studies (τ^2) relative to total variation, ($\tau^2 + \hat{se}_j^2$). More specifically, the degree of pooling λ is given by:

$$\lambda = 1 - \frac{\tau^2}{\tau^2 + E(\hat{se}_j^2)}$$

where $\lambda = 1$ corresponds to the full pooling case, and $\lambda = 0$ corresponds to the no pooling case.

In Table II.5, I provide the pooling metrics estimated for each of the studies in the baseline sample. For all but one study, I find that the pooling metric is greater than 0.5, suggesting that study-level estimates are being adjusted towards the population mean. The overall pooling factor across studies suggests that 89% of the heterogeneity in estimated effects is due to sampling variation. Thus, genuine heterogeneity across studies is

[6]Note here that the Vanilla subset would not be comparable to Engel (2011), since he includes all dictator games in his sample that report gender differences in giving (and not just one-shot, standard dictator games).

low and each additional study is informative on the overall population effect.

The intuition of this result can also be seen graphically, in Figure II.3. Here, I plot the posterior effect estimates of each study in my baseline sample, and the corresponding 95% probability intervals for a model with full pooling, partial pooling, and with no pooling. Compared to full pooling, the 95% probability intervals of the partial pooling model are larger, capturing the fact that there is some heterogeneity across studies. These bounds are much smaller than that of the original study estimates, however, suggesting that differences in effects across studies are primarily driven by sampling variation, rather than genuine variation.

II.5 Publication Bias

The results from the Bayesian hierarchical model can be interpreted as the overall gender differences in giving for dictator games, within settings for which researchers have conducted dictator games and critically, within the population of working papers and journals that publish results on gender differences in dictator game giving. The extent to which the result can be applied to our broader understanding of gender differences in altruism thus depends on the external validity of behaviour in dictator games (and lab experiments more generally), and the degree of publication bias. While there is extensive literature on the former issue (e.g. List, 2007; Levitt and List, 2007; Benz and Meier, 2008; Franzen and Pointner, 2013), I now turn my focus to exploring the extent of publication bias.

In particular, the findings from the Bayesian hierarchical model would potentially be biased in the presence of publication bias, that is, if certain types of results are systematically more likely to be published. Importantly, in exploring 'publication bias' I am unable to distinguish between the decisions of the journal and the decisions of the researcher, otherwise known as the 'file drawer' problem (Rosenthal, 1979). The issue of the 'file drawer' problem is particularly relevant to this setting, since almost all studies of the dictator game collect data on gender, but only a select subset report the average giving of women versus men. In fact, authors of 6 studies surveyed in my data collection state explicitly that they do not find statistically significant gender differences in giving, and hence do not report the results. Are researchers more likely to report gender results if they find large differences in giving? Similarly, are editors more likely to publish results that find a large effect? In this section, I explore the extent to which this may be true.

I start the section by documenting the patterns in the distribution and variation in estimated treatment effects and standard errors across studies. Then, I follow Andrews and Kasy (2019) in estimating the conditional probability of publication using a truncated selectivity model. Under the assumption that the latent variables are independently and identically distributed, the model allows me to parametrically estimate how the probabil-

ity of publication varies with the study results. Finally, I present the results and discuss the implications of this analysis.

II.5.1 Distribution of estimates

As a first pass for exploring the degree of publication bias, it is useful to consider the distribution of test statistics, point estimates, and standard errors of the full set of experiments. I follow Andrews and Kasy (2019) and Brodeur et al. (2016) in considering the distribution of the z-statistics (the ratio of the effect size to the standard error) above and below the 5% significance level threshold. Intuitively, absent publication bias, there should not be any bunching or jumps in the test statistics on either side of the significance thresholds.

I focus here on three subsets of the data that may be of interest. First, the *FullData*, comprising of all 100 experiments in my sample. Second, the *GenderTopic* subset, comprising of the 65 experiments from the 20 studies that explicitly refer to a gender-related term in the title of the paper. Third, the *TopJIF* subset, comprising of the 57 experiments in my full sample from the 19 studies published in top peer-reviewed journals. As a proxy measure for journal quality, I use the Journal Impact Factors (JIF) published in the 2019 Journal Citation Reports, which give a measure of the impact and influence of an academic journal. I include in my *TopJIF* the subset of results from papers published in peer-reviewed journals that are ranked in the top two quartiles of the 5 year average JIF indicators [7] .

In Figure II.4, I construct a binned-density plot of the z-statistic for the full dataset, the *GenderTopic* subset, and the *TopJIF* subset. Similar to Brodeur et al. (2016) I observe jumps in the distribution around the cutoffs for -1.96, 0, and 1.96 for the full dataset. This pattern is broadly similar for the subsets with slight differences: while for the *GenderTopic* subset, there does not appear to be a jump in the data around 1.96; for the *TopJIF* subset,

[7] Sutter and Kocher (2001) find that the JIF rankings in economics remain stable over time: 95% of economics journals remain in the same or neighbouring quartile over a 10-year period; and there is even less variation in JIF for the Top 15 journals. Thus, although papers in my sample are published at different times, the 5-year average JIF, is likely to be a good proxy for journal quality at the time of publication.

the jump in the density is noticeably smaller around zero.

Next, I construct funnel plots of the effect estimate against the standard errors in Figure II.5, as suggested by Andrews and Kasy (2019). Absent publication bias, as the standard error of a study increases, the effect estimates should get noisier and be symmetrically split to the right and left of the true effect. As with the density plots, any bunching around the significance thresholds (as illustrated by the dotted lines) would again be suggestive of some degree of selective publication. As seen in Figure II.5 there is a mass of effect estimates asymmetrically bunched around positive effect sizes that are statistically distinguishable from zero at the 5% level. This is seen for all three subsets, but particularly evident for the full sample, as seen in panel A.

II.5.2 Methodology

I follow Andrews and Kasy (2019) in modelling publication bias as a truncated sampling process, in which studies are selected for publication only on the basis of the results. Let us distinguish between latent (unobserved) variables, denoted by an asterisk (e.g. $\hat{y}^*_{jk}, y^*_{jk}$), which capture the full set of experimental results; and observed variables (e.g. $\hat{y}_{jk}, y_{jk}$), which capture the subset of the latent results that are published in journals or in working papers. In particular, we observe $\hat{y}^*_{jk}$ only if $D_{jk} = 1$, that is, if the result is published.

Assume $(\hat{y}^*_{jk}, y^*_{jk}, \hat{se^2}^*_{jk}, D_{jk})$ are jointly iid across j and k with

$$\hat{y}^*_{jk} \sim N(y^*_{jk}, \hat{se^2}^*_{jk})$$

$$y^*_{jk} \sim N(\mu^*, \tau^{2^*})$$

$$D_{jk} \mid \hat{y}^*_{jk}, y^*_{jk}, \mu^* \sim Ber(p(Z^*))$$

$$\text{where } \hat{y}_{jk} = \begin{cases} \hat{y}^*_{jk} & \text{if } D_{jk} = 1 \\ unobserved & \text{if } D_{jk} = 0 \end{cases}$$

$$\text{and } p(\hat{y}_{jk}/\hat{se}_{jk}) \propto \begin{cases} \beta_{p,1} & \hat{y}_{jk}/\hat{se}_{jk} < -1.96 \\ \beta_{p,2} & -1.96 \leq \hat{y}_{jk}/\hat{se}_{jk} < 0 \\ \beta_{p,3} & 0 \leq \hat{y}_{jk}/\hat{se}_{jk} < 1.96 \\ 1 & \hat{y}_{jk}/\hat{se}_{jk} \geq 1.96 \end{cases}$$

In the above model, an experiment is published with probability $p(Z)$, where Z is the z-statistic, calculated as the ratio of the estimated treatment effect and corresponding standard error. I assume that the probability of publication differs by the intervals of the test statistic around the 5% significant level (where the null hypothesis is a zero effect size), and allow for asymmetric selection depending on the sign of the estimated result.

Relative to experiments that find a positive result that is significant and distinguishable from zero at the 5% level, positive and insignificant results are $\beta_{p,3}$ as likely to be published, negative and insignificant results are $\beta_{p,2}$ as likely to be published, and finally, negative and significant results are $\beta_{p,1}$ as likely to be published.

Under the assumption that the latent variables are independently and identically distributed, Andrews and Kasy (2019) show that we can parametrically identify and estimate $p(z)$ up to scale. Note here that while the independence of latent variables cannot be tested by construction (since we do not observe studies that are not published), a clear violation in this setting is the fact that I observe results from multiple experiments within the same study. To account for this, I assume conditional independence and cluster standard errors by study j. In the proceeding section, I estimate the conditional probability of publication, $p(z)$, using the maximum likelihood estimation set out in Andrews and Kasy (2019).

II.5.3 Results

I estimate the conditional probability of publication across the three sub-samples of my data: the Full dataset, the *GenderTopic* subset; and the *TopJIF* subset. The results are presented in Table II.6.

Using the full sample of experiments, I find strong evidence of selection based on statistical significance. As seen in columns 4-6 of Table II.6, positive results (where women give more than men) that are statistically distinguishable from zero at the 5% level are over 13 times more likely to be published than statistically significant negative results that find that men give more than women; and over 3 times more likely to be published than results that are negative and statistically insignificant. While the magnitude of $\beta_{p,3}$ suggests that results that are positive and statistically significant are more likely to be published than those that are positive and statistically insignificant, this difference is not significant at conventional levels.

Restricting the full sample of experiments now to papers that explicitly study gen-

der, the *GenderTopic* subset, I find evidence for selection based on statistical significance, but not on the sign. Among dictator games that study gender differences, experiments that find a statistically significant and positive result are over 4 times more likely to be published than a negative and insignificant result, and over 2 times more likely to be published than a positive and insignificant result.

Selection based on statistical significance is less severe in the $TopJIF$ subset, the sample of results from peer-reviewed journals with the highest journal impact factors. Positive and significant results are around three times more likely to be published than positive and non-significant results. Compared with the two other subsets, the relative probability of publishing negative results (significant or insignificant) is higher compared with the two other subsets. In fact, the magnitude of $\beta_{p,1}$ suggests that negative and significant results are more likely to be published than positive and significant results, although this difference is not statistically significant at conventional frequentist levels.

II.5.4 Implications for Bayesian inference

What do these results mean for Bayesian inference? The implications for posterior inference depend on the distributional assumptions on the hyperparameters, μ and τ. Andrews and Kasy (2019) distinguish between two extreme classes of priors: unrelated parameters and common parameters[8]. Whereas under unrelated parameter priors, posterior inference is unaffected by publication bias, under common parameters priors, inference is affected and the posterior distribution would need to be adjusted using the truncated likelihood. Similarly, Yekutieli (2012) show that under 'fixed', non-informative priors, Bayesian inference needs to be adjusted for selection.

In the context of this study the hyperpriors, μ and τ, likely lie between the two extremes of unrelated and common parameters. Hence posterior inference from the two-stage model is likely to be affected by selection.

[8] Andrews and Kasy (2019) define unrelated priors as the case in which the prior distribution is a point mass around a value; whereas common parameters priors, are such that the prior distribution assigns positive probability to point-measures of the prior.

Ideally, I would quantitatively adjust the posterior effect estimates to account for selective publication. As seen in Section II.5 however, the form of selection bias appears to operate in a complex way, and the conditional probability of publication differs by both the topic of the study and by the quality of the journal. Hence a blanket uniform adjustment of the posterior treatment effect is unlikely to be forthcoming.

Taken together, these results suggest that the estimate from the Bayesian hierarchical model is likely to provide an upper bound estimate of the overall effect for the wider population.

II.6 Conclusion

By aggregating results from dictator game experiments, I make two key contributions. First, I estimate the average gender difference in dictator game giving using a Bayesian hierarchical model that allows me to separate between sampling variation and genuine heterogeneity across studies. Second, I contribute to the interpretation of these studies, by exploring how the prevalence of publication bias affects the results available in published and working papers.

I find that given the available evidence, women give 3 percentage points more than men in dictator games. This effect is smaller than that found in the most frequently cited studies, and the estimated 95% probability interval of 1.7 to 4.5 percentage points rules out existing estimates of the aggregate gender effect (e.g. Engel, 2011). I show that the observed gender differences are likely driven by publication bias, whereby papers are selected based on statistical significance. Thus, while the average giving of women relative to men is 3 percentage points among published results, the true effect for the wider population is likely to be smaller.

Given that lab experiments routinely collect data on gender (but may or may not report the findings), my results also highlight the importance of standardized reporting and data transparency to facilitate comparability across studies.

While previous research argues that gender differences in dictator game giving are driven by experimental design, I show that even in the presence of contextual differences, estimates of gender differences in altruism are likely to overestimate the effect due to selective publication. Although I do not explicitly study the role of experimental characteristics in this paper, understanding the relative importance of publication bias versus experimental design would be an interesting direction for future research.

II.7 Tables and figures

II.7.1 Figures

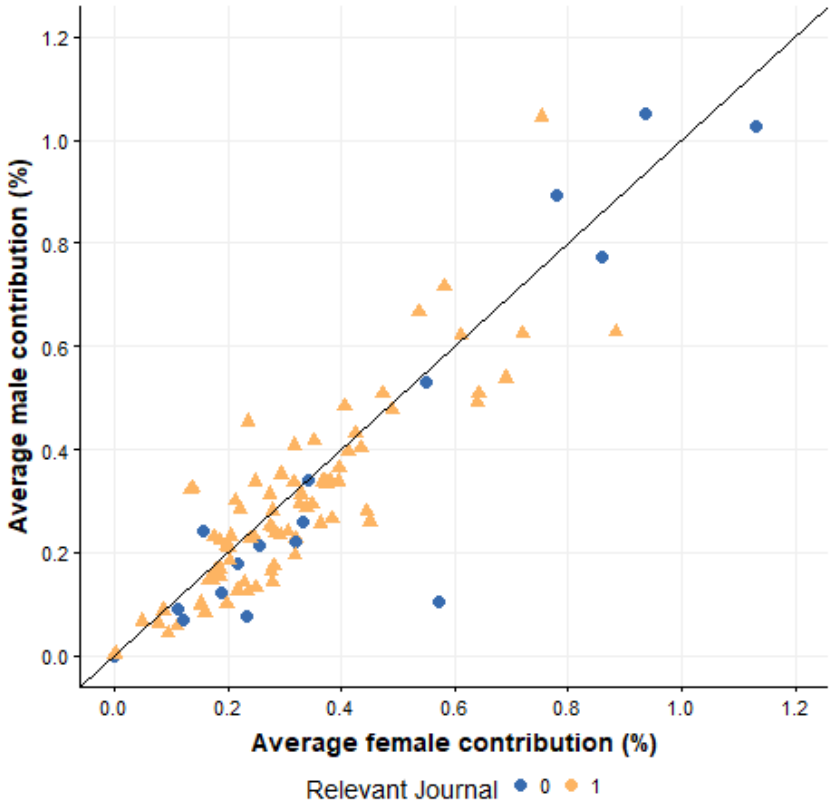


FIGURE II.1: Average contributions of women versus men (% of stake size), by journal type.

Notes: Relevant Journals defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). See Footnote 2 for full list.

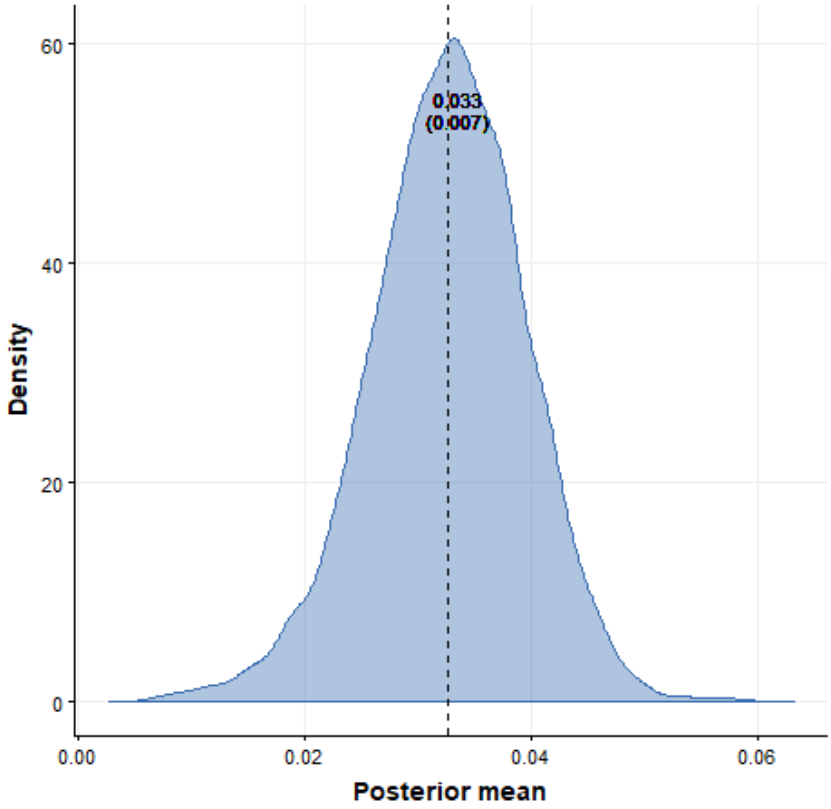


FIGURE II.2: Posterior distribution of effect estimate

Notes: Posterior distribution of the estimated of the gender difference in contributions, defined as the percentage point difference in contribution of women relative to men.

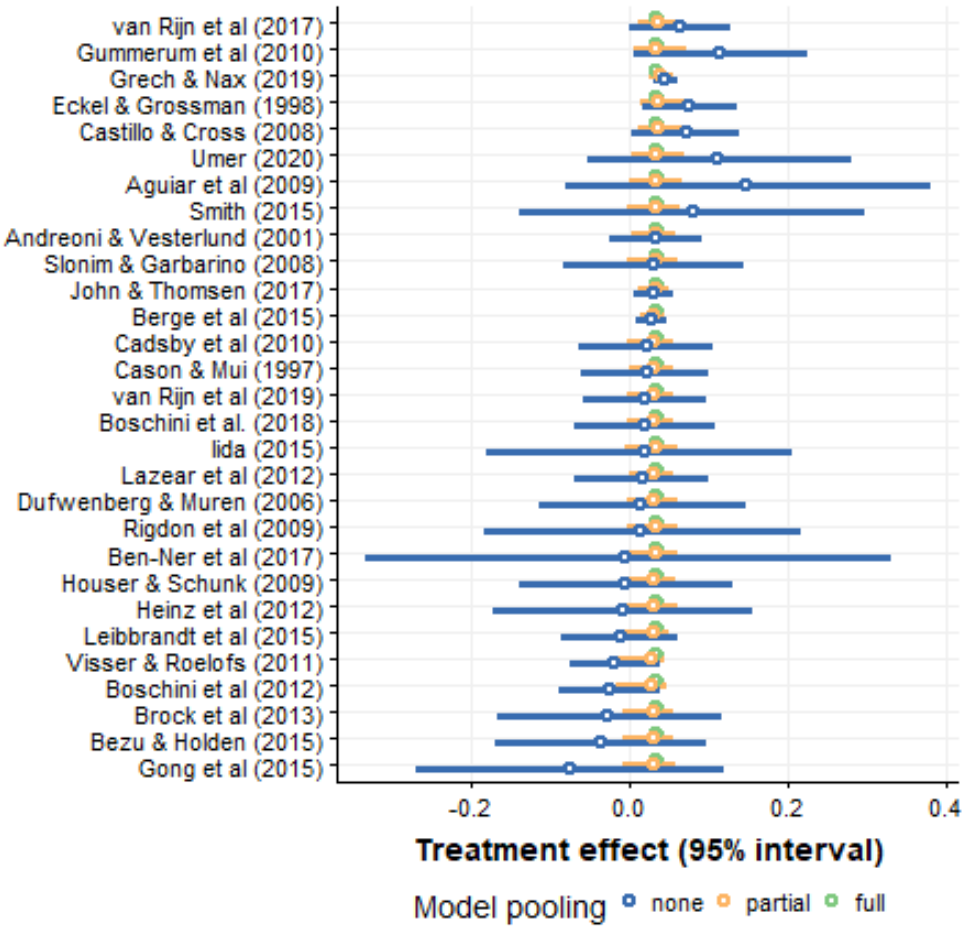


FIGURE II.3: Model comparison - posterior effect estimates of μ by study

Notes: Estimated posterior mean and 95% probability intervals across models of zero, partial, and full pooling.

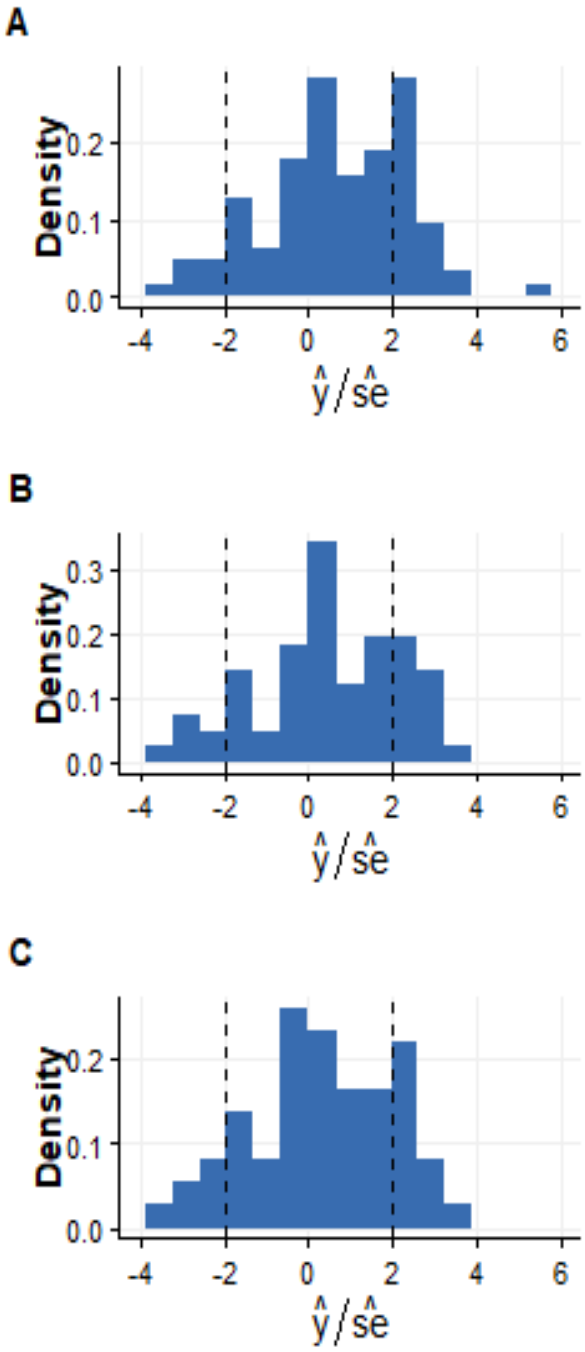


FIGURE II.4: Binned density plot of estimated z-statistics

Notes: The dotted lines mark where the z-statistic is equal to 1.96 and -1.96. Panel A: full dataset; Panel B: Gender topic subset, including only observations with 'gender' in the title; Panel C: Top JIF subset, including only observations published in peer-reviewed journals placed in Q1 & Q2 of 5 year Impact Factor rankings

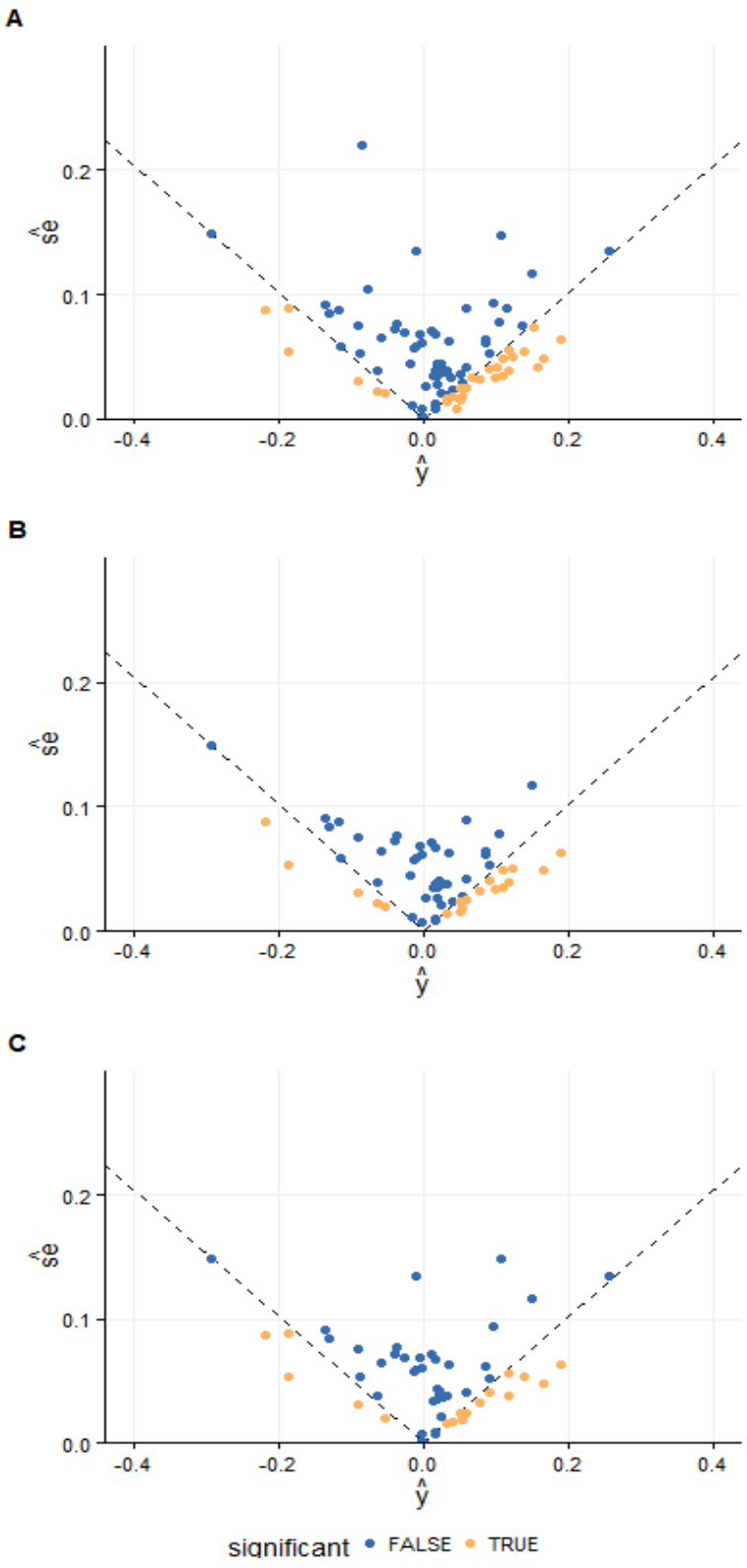


FIGURE II.5: Funnel plots of effect estimates

Notes: Scatter plot of effect estimate and standard error, by whether or not observation is statistically distinguishable from zero, at the 5% level. The dotted lines mark where $|\hat{y}/\hat{se}| = 1.96$. Panel A: full dataset; Panel B: *GenderTopic* subset, including only observations with 'gender' (or gender related term) in the title; Panel C: *TopJIF* subset, including only observations published in peer-reviewed journals placed in Q1 & Q2 of 5 year JIF ranking

II.7.2 Tables

TABLE II.1: Selection Criteria

	Selection Criteria	Description
1	Keywords	Dictator Game AND altruism, philanthropy, generosity, or intergenerational transfers
2	Experimental Design	Focus on non-interactive, single stage dictator games. Exclude sequential or multidimensional dictator games; games which give recipient power; games which require effort to generate the endowment
3	Results Reported	Results on either (1) Average contributions of men and women or the gender differences in contributions, and the associated standard errors, or (2) Raw data to calculate.

TABLE II.2: Average contributions by gender, % stake size

	N	Mean	St. Dev.	Min	Max
Average contribution of men	100	0.297	0.217	0.000	1.052
Average contribution of women	100	0.324	0.208	0.000	1.131
Gender difference in contribution	100	0.027	0.097	−0.292	0.465

Notes: Gender difference in contribution defined as the percentage point difference in contribution of women relative to men. Positive gender difference corresponds to women giving more than men. A contribution of more than 1 corresponds to experiments in which the price of giving is less than 1 (see: Andreoni and Vesterlund, 2001, for an example).

TABLE II.3: Average contributions by gender, % stake size - baseline sample

	N	Mean	St. Dev.	Min	Max
Average contribution of men	69	0.289	0.185	0.001	1.045
Average contribution of women	69	0.303	0.175	0.00001	0.883
Gender difference in contribution	69	0.013	0.086	−0.292	0.257

Notes: Gender difference in contribution defined as the percentage point difference in contribution of women relative to men. Positive gender difference corresponds to women giving more than men. A contribution of more than 1 corresponds to experiments in which the price of giving is less than 1 (see: Andreoni and Vesterlund, 2001, for an example). Baseline sample includes all experiments published in *RelevantJournals*, other than those that include priming and framing. *RelevantJournals* are defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). See Footnote 2 for full list.

TABLE II.4: Posterior estimates of μ, by subsample

					Quantiles				
	N	J	Mean	s.e.	2.5%	25%	50%	75%	97.5%
Baseline	69	29	0.0323	0.0069	0.0171	0.0284	0.0328	0.0370	0.0454
Full sample	100	38	0.0319	0.0046	0.0229	0.0289	0.0319	0.0351	0.0406
Vanilla	31	23	0.0441	0.0120	0.0197	0.0369	0.0443	0.0514	0.0690

Notes: Baseline sample: includes all experiments, other than those that include priming and framing, and that are published in *RelevantJournals*. *RelevantJournals* are defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). Full sample: includes all experiments and results. Vanilla sample: includes only standard, one-shot dictator games, published in *RelevantJournals*.

TABLE II.5: Pooling factors for each study

	Pooling factor
Overall Pooling	0.892
Aguiar et al. (2009)	0.986
Andreoni and Vesterlund (2001)	0.863
Ben-Ner et al. (2017)	0.993
Berge et al. (2015)	0.550
Bezu and Holden (2015)	0.963
Boschini et al. (2012)	0.876
Boschini et al. (2018)	0.924
Brock et al. (2013)	0.963
Cadsby et al. (2010)	0.911
Cason and Mui (1997)	0.916
Castillo and Cross (2008)	0.872
Dufwenberg and Muren (2006)	0.960
Eckel and Grossman (1998)	0.868
Gong et al. (2015)	0.981
Grech and Nax (2020)	0.489
Gummerum et al. (2010)	0.946
Heinz et al. (2012)	0.974
Houser and Schunk (2009)	0.962
Iida (2015)	0.981
John and Thomsen (2017)	0.640
Lazear et al. (2012)	0.919
Leibbrandt et al. (2015)	0.896
Rigdon et al. (2009)	0.982
Slonim and Garbarino (2008)	0.946
Smith (2015)	0.986
Umer (2020)	0.975
van Rijn et al. (2017)	0.871
van Rijn et al. (2019)	0.906
Visser and Roelofs (2011)	0.846

Notes: Pooling factors correspond to the metric suggested by Gelman and Pardoe (2006). The overall pooling factor is an arithmetic mean of the pooling factors across studies.

TABLE II.6: Estimates of $p(z)$, by subset

			$p(z)$			Interpretation[1]		
	Studies	N	(1) $\beta_{p,1}$	(2) $\beta_{p,2}$	(3) $\beta_{p,3}$	(4) $1/\beta_{p,1}$	(5) $1/\beta_{p,2}$	(6) $1/\beta_{p,3}$
Full sample	38	100	0.074	0.312	0.705	13.514	3.205	1.418
			(0.131)	(0.298)	(0.529)			
GenderTopic	20	65	0.245	0.224	0.403	4.082	4.464	2.481
			(0.456)	(0.193)	(0.232)			
TopJIF	19	57	1.543	0.423	0.333	0.648	2.364	3.003
			(1.077)	(0.334)	(0.241)			

Notes: (1) A positive and significant result is $1/\beta_{p,1}$ more likely to be published than a negative and significant result; $1/\beta_{p,2}$ more likely to be published than a negative and insignificant result; and $1/\beta_{p,2}$ more likely to be published than a positive and insignificant result. (2) *GenderTopic* subset, includes only observations with 'gender' (or a gender related term) in the title; *TopJIF* subset, includes only observations published in peer-reviewed journals placed in Q1 & Q2 of the 5 year Journal Impact Factor rankings in the 2019 Journal Citation Reports.

Chapter III

Men are from Mars, and Women Too: a Bayesian Meta-analysis of Overconfidence Experiments

Joint with: Oriana Bandiera, Nidhi Parekh, and Barbara Petrongolo

Abstract: Gender differences in self-confidence could explain women's under representation in high-income occupations and glass-ceiling effects. We draw lessons from the economic literature via a survey of experts and a Bayesian hierarchical model that aggregates experimental findings over the last twenty years. The experts' survey indicates beliefs that men are overconfident and women under-confident. Yet, the literature reveals that both men and women are typically overconfident. Moreover, the model cannot reject the hypothesis that gender differences in self-confidence are equal to zero. In addition, the estimated pooling factor is low, implying that each study contains little information over a common phenomenon. The discordance can be reconciled if the experts overestimate the pooling factor or have priors that are biased and precise.

III.1 Introduction

Gender inequality in the labour market is rife. Women make systematically different education choices from men, are under-represented in high-earning careers, and they bear the bulk of the earning penalty associated to parenthood (Kleven et al. 2019).

There are two, fundamentally different, explanations for this difference. The first is that men and women are equal in all relevant dimensions but face different opportunities or constraints. In this case, gender inequality can be a symptom of misallocation and

policies that promote gender equality can increase efficiency. The second is that men and women have different psychological traits that drive educational choices and labour market outcomes. In this case, gender inequality in labour outcomes is a manifestation of gender differences in traits. We contribute to this debate by aggregating the evidence on gender differences in traits, with a particular focus on overconfidence.

Several lab, and more recently field, experiments (surveyed, among others, by Croson and Gneezy 2009b, Bertrand 2018, 2011a, Azmat and Petrongolo 2014) have investigated gender differences in personality traits. Those that matter for labor outcomes can be grouped in three broad areas: attitudes towards risk, social preferences and confidence. These shape decisions at every stage of a person's career, from years of schooling to major choices, from job applications to the choice of sector and firms, and, once at work, on pay and promotions.[1] Knowing whether there are systematic gender differences in these traits is key to interpreting differences in outcomes but drawing definitive conclusions from the large body of experiments is limited by the fact that due to differences in settings, stakes and design, findings cannot be easily aggregated.

This is what we attempt to do: aggregate findings. We begin by surveying experts - academic economists - who are the main "consumers" of this literature and draw lessons from it. Our sampling frame is the universe of research fellows and affiliates of CEPR - 1300 economists based mostly in Europe and the US. Our sample of respondents (342, a 26% response rate) are asked to score men and women on risk aversion, confidence and altruism in a 0-100 scale, based on their reading of the literature. Men come out more confident, risk loving and selfish. The mode of the gender gap is however close to zero for risk and altruism, as experts who rank men highly also rank women highly on the same dimension. In contrast, the modal gap for overconfidence is positive and the within person correlation is negative, as experts who rank men highly overconfident

[1]High-income careers typically develop in competitive environments, in which winners may be disproportionately rewarded, and are characterized by a relatively high variability of earnings. Individuals who are unwilling to compete or are particularly risk-averse may thus simply not embark on those careers. Likewise, pro-social preferences may lead to choices that do not maximize own monetary payoffs. These are only some of the channels whereby gender differences in these traits may interfere with women's labour market success.

typically rank women as very underconfident. On average, men are rated overconfident and women underconfident.

In the second part of the study we assemble a data set made of all the experimental tests of confidence published in the last 20 years and estimate a Bayesian Hierarchical Model to aggregate the findings. The model tackles the main challenge of aggregating evidence from different settings and experimental designs head on by estimating, together with the gap, a "pooling factor" which measures the extent to which each result is informative about a common phenomenon versus its own context-specific effects.

Why confidence? Self-confidence is important for understanding selection into certain education tracks, occupations, and careers, as well as ex-post payoffs to these choices, which depend on group composition whenever one's performance is remunerated against the performance of peers. When remuneration has a zero-sum, it pays off to be realistic about one's chances of success. In other words, underconfident individuals compete too little and overconfident individuals compete too much relative to the choices that would maximise their expected payoffs. But in real-life circumstances remuneration rarely has zero-sum, as the expected return from several competitive settings – such as the expected outcome of a job application or a promotion – is positive. This is because the worst-case scenario is typically one's status quo. By shying away from such situations, underconfident individuals forgo positive chances of success. In addition, they may also miss out on feedback and experience that could be gained from participating, regardless of outcomes. Hence it is important to establish whether some groups are likely to be held back in the labour market by a tendency to underestimate their performance, whether absolute or relative to peers. For example, in the study by Niederle and Vesterlund (2007), men and women overstate own performance in a simple arithmetic task at which they are on average equally good, but men tend to be more overconfident about their performance than women. This gap in overconfidence explains part of the gender difference in the choice of tournament over piece-rate compensation for the same task.

Our sample includes papers that appear between 2000 to 2020 in peer-reviewed economic journals or widely circulating working paper series, that provide a measure of

confidence for men and women on a real task, whether in lab or field settings. We identify 38 papers that satisfy these criteria, providing a pool of 90 paired results. In stark contrast with the experts' assessment, 72% of the estimates indicate that both genders are overconfident, while only 18% are consistent with the shared belief that men are overconfident and women and underconfident.

Our Bayesian analysis can be performed on a subsample of 39 experiments that provide standard errors for their adopted overconfidence metric. Our main results deliver a difference in the (standardised) overconfidence of men and women of 0.094, with a standard error of 0.054. Overall, men are more confident than women, but we cannot reject that this difference is zero at the 5% level. This is in line with the findings of BHM estimates of gender gaps in altruism (Rao, 2020) and response to incentives (Bandiera et al., 2021). In contrast to these studies, however, the pooling factor is low and the relatively large posterior variance implies that each individual study is poorly informative about a common phenomenon. Based on a standard pooling metric used in the Bayesian hierarchical literature (Rubin, 1981; Gelman et al., 2013), we estimate that only 23% of variation across studies is sampling variation, with more than three quarters of the variation being driven by genuine differences across studies.[2] This implies that, if we were to run a new experiment, we could not be reasonably confident that the resulting gender difference in overconfidence would be close to the posterior mean in the original sample. A naive meta analysis that ignores this heterogeneity and assumes that each study contributes to identify a common effect yields an estimate of 0.114 (s.e. 0.013).

The full-pooling model can (qualitatively) match the profession's beliefs about gender gaps in overconfidence, but it cannot explain why the majority of economists believe that women are underconfident.

We consider two further explanations within a general BHM with partial pooling. First, we acknowledge that different results may achieve varying degrees of visibility in the profession, implying different intensities of belief updating. We hypothesise that

[2]One possible interpretation is that the context matters more than individual preferences for confidence experiments, i.e. the same individual might be over or under confident depending on the circumstances.

the salience of various results is proportional to the citations they receive, which are in turn related to how long they have been in the public domain, the prestige of the outlet where they are published and the authors' prominence. We estimate a modified BHM in which the precision of the study estimates is adjusted by the citations received, and obtain posterior estimates that are very close to those from the original model, with a posterior mean gap in confidence of 0.080, and a standard error of 0.068. The conclusion is that results delivering a larger gender gap in confidence do not systematically obtain more cites, hence the distribution of citations across papers may not explain the observed beliefs among economists.

Second, we explore the role of prior beliefs in shaping the updating process. Our BHM assumes a normally distributed prior for the average gender difference in overconfidence, with mean zero and unit variance. This reflects that we are a priori agnostic about which gender is more overconfident, but we allow for substantial variation around such zero mean. We next consider a biased prior (for example one that is commensurate with gender differences in overconfidence observed in survey responses), while keeping a unit variance around it, and obtain nearly identical posterior estimates to those from the original BHM with unbiased priors. Only when we significantly increase the precision of the biased prior, do we obtain a positive and precise posterior estimate of the gender gap in overconfidence. In other words, when priors are very precise, they are hardly updated when new information is received. If the prior is biased and very precise, so will be the posterior.

In summary, we conjecture that biased beliefs in the profession could stem from extreme priors or lack of Bayesian updating. The full-pooling model described above is one special case of updating, which delivers much more precise results than the general BHM, but is not supported by the data. Other cases could involve selective or non-probabilistic updating, such that individuals interpret information to confirm what they believe in the first place, or form beliefs that are deterministic functions of their information sets (see for example Jackson et al. 2021). Lack of Bayesian updating could be rationalized in terms of its cognitive costs, whereby simplifying the updating process is a way to save

on cognitive effort. As a consequence, gender stereotypes may arise as generalisations that help individuals economise on cognitive resources when forming beliefs about the characteristics of groups members.[3].

Our findings are in line with the social psychology literature that benchmarks between-gender differences in traits to within-gender differences. The idea is to focus not just on the difference in means between men's and women's characteristics, but also take into account the overlap between the respective distributions. The findings in this literature indicate that, for most relevant traits related to cognitive and non-cognitive abilities – including self-confidence – within-gender differences are much larger than between-gender differences, so there is a substantial overlap in the gender-specific distributions (Hyde 2005, 2014, Bertrand 2020).

The rest of the chapter is organised as follows. Section 2 describes the survey of economists' beliefs. Section 3 describes the data including the process of study selection and the Bayesian sample. Section 4 discusses the empirical approach. The results are described in Section 5. Section 6 highlights the knowledge gap between expert's beliefs and the evidence, and considers model extensions that could explain the differences between our meta analysis and survey results. Section 7 concludes.

III.2 Experts' survey

In the fall of 2019, we surveyed the universe of Research Fellows and Affiliates of the Centre for Economic Policy Research (CEPR), a leading research network in Europe. The survey was sent by email (see Figures C1 and C2 in the Appendix for a screenshot of the full survey) to 1,300 economists and the response rate was 26%, i.e. 342 experts responded. Respondents were asked to rate men and women on three traits -confidence, risk aversion and altruism- *based on their reading of the literature*.

Ratings are given on a scale of 0 to 100. We set 50 as the neutral point in the risk

[3]See the related discussion in Bertrand (2020). See also Bordalo et al. (2019) for evidence on the role of stereotypes in shaping beliefs about gender skills.

question, realistic for the confidence question, and care equally about self and others for the altruism question. Of the 342 researchers who responded, 64% were male, 57% were full professors; Applied Micro is the modal field, accounting for 32% of the observations. Table C1 in the Appendix reports a full breakdown of the respondents' field of specialization.

Figure III.1 shows the mean answers as well as the gender gap (men minus women) on all three measures. The figure shows positive gaps on all three dimensions – i.e. men are more confident, more risk loving, more selfish – but the confidence gap is much larger as men are rated twice as confident as women and the average expert rates women as underconfident (35 out of 100) and men as overconfident (69 out of 100), with 77% of the surveyed population rating women as underconfident and men and overconfident.

Figure III.2 shows the scatter plots of male and female scores on the three dimensions. Each dot represents the grades given by one person. We find a positive correlation for both altruism and risk, that is experts who score men high also score women high. In contrast, the correlation for confidence is negative, that is experts who score men high also score women low. This partly explains why the gap is so much larger.

Figure III.3 shows the distributions of the three gaps. The confidence gap stands out both because there is much more variation across experts and because the mode is positive, while it is zero for the other two. Figures C3, C4 and C5 in the Appendix shows scatterplots of experts' responses on the confidence question by gender, field of specialization and seniority, and reveals no evidence of systematic differences in responses along any of these dimensions.

Interestingly, respondents who report that men and women are very similar on altruism (within ten points of one another) estimate that men are more confident than women by 23 points. And respondents who report that men and women are very similar on risk estimate that men are more confident than women by 22 points. Thus, even respondents who believe that men and women are fairly similar on other traits – a belief in line with in the meta-analyses of Bandiera et al. (2021) and Rao (2020) – believe that they differ in

over-confidence. The next section will show how findings from the experimental literature compare to economists' beliefs about gender differences in over-confidence.

III.3 Data

III.3.1 Paper selection and summary evidence

We select papers that appeared in the public domain during the past two decades, a very active period for the literature on gender differences in psychological traits. It is important to note that selection into this sample biases our estimate against the null of zero effects as publication is on average biased in favour of significant results (Kasy 2021). By searching google scholar and RePEC online repositories for papers with keywords "confidence" and "gender", we identified 474 such papers that appeared during 2000-2020. We next selected those published in the top 100 economic journals (according to RePEC), or in the NBER, CEPR, or IZA Working Paper series, yielding 140 papers. We finally checked each of these papers for relevant information on confidence by gender, selecting papers in which confidence was measured in relation to actual performance in a specific task. Specifically, we scouted for information on either:

(a) The difference between self-assessed performance score or rank (according to the specific study setting), and the respective actual performance score or rank, by gender.[4] Based on this difference, we obtain the average degree (or intensity) of overconfidence by gender for each study. This is also referred to as the "intensive margin" measure.[5]

(b) The difference in the share of men and women who are overconfident or underconfident. This is referred to as the "extensive margin" measure. It is calculated based

[4]Measures based on performance scores provide an estimate of absolute overconfidence, while measures based on performance rank in a tournament provide an estimate of relative overconfidence, both of which are used in this paper.

[5]For example, Kamas and Preston (2012) compare participants' actual score to their estimated score in a math task to measure their self-confidence. In another experiment, they compare participants' actual ranking in a group competition to their estimated ranking.

on one of the two methods below: (i) the share of men and women who overstate (understate) their performance score (rank);[6] (ii) the share of men and women who self-select into a tournament, believing that they will win the tournament, but do not, or the share of men and women who do not self-select into a tournament, believing that they will not win, but would have won based on their performance. [7]

Our working sample consists of 90 studies, i.e. paired observations on self confidence for men and women, from 38 papers that meet the criteria laid out previously. The list of papers is provided in Table C2 in the Appendix. 71 observations are obtained in the lab, 9 in the field, and 10 from combinations of lab and field experiments. Most experiments are based on a student subject population in a high-income country.

Based on this sample we build a dataset containing relevant measures of overconfidence by gender with the associated metric of statistical significance (whenever available) for each experiment included in the papers, as well as information on authors, publication outlet and impact factor. For working papers, the journal impact factor is imputed assuming the paper will be eventually published in the journal where the most cited author is most published.

III.3.2 Bayesian analysis sample

To aggregate evidence across studies using Bayesian hierarchical methods, we need estimates of the standard errors for each result included in the analysis. We therefore further

[6] For example, in Reuben et al. (2017), students perform addition tasks under tournament and piece-rate compensation. After completing the tasks, students are asked to rank their beliefs on their performance within a group of four. The authors measure confidence by comparing the percentage of men and women who think they would rank first versus the percentage who would have come first, based on their performance.

[7] For example, Dreber et al. (2014) consider the share of boys and girls who choose to compete in a verbal task tournament. As participants are only compensated if they win a tournament, the participation rate is used to measure one's beliefs of outperforming other participants. The authors find that 33% of boys choose to compete in the verbal task, compared to 28% of girls. However, based on performance in the tournament, the probability of winning is similar for boys and girls, implying that as many girls as boys should have chosen to compete. In this context, the authors measure confidence as the difference between the share of boys and girls who choose to compete versus those who *should* compete, as proxied by their actual performance.

select results for which standard errors are reported (or could be obtained from reported p-values or t-statistics) for the gender gap in overconfidence. 39 studies (from 16 papers) meet these criteria. Among these, 24 studies also report standard errors separately for male and female overconfidence.

Figure III.4 compares the distribution of the raw results across the different samples. The first bar refers to the whole sample of papers that provide paired observations on self confidence for men and women; the second bar refers to the subsample that also provides confidence levels for the gender difference in over-confidence; and the third bar refers to the subsample that provides significance levels separately for men and women. Irrespective of selection criteria, the vast majority of studies find that both men and women are over-confident, while only a minority finds that men are overconfident and women are underconfident. The subsamples on which we perform the Bayesian analysis are therefore representative of the larger population of papers that measure gender differences in overconfidence.

Further details on the analysis samples are reported in Table III.1. 26 out of 39 studies provide measures of the degree of overconfidence among men and women (the intensive margin sample). Of these, 17 studies report standard errors separately for men and women. Men and women on average overestimate their score (or underestimate their rank) by 4 and 2.7 percentage points respectively, and the average gender gap is 2.9 percentage points. The remaining 13 studies only report shares of overconfident men and women (the extensive margin sample). Of these, only 7 report standard errors separately for men and women. The data reported imply that 52.2% of men and 46.2% of women overestimate their performance, respectively, and on average the share of men overestimating their performance exceeds the female share by 8.5 percentage points. Measures of overconfidence from the two subsamples can be combined in a standardized measure, given by the specific metric, divided by the within-sample standard deviation (Cohen 2013). Men and women overestimate their ability relative to their performance by 0.421 and 0.323 standard deviations, respectively; and men overestimate their ability by 0.115 standard deviations more than women.

In the empirical analysis that follows we will provide posterior overconfidence estimates for each sub-sample, as well as the full sample, for men and women separately and for the corresponding gender gap.

III.4 Empirical Approach

By combining information from several data sources potentially interrelated, meta-analysis naturally lends itself to hierarchical modelling. The Bayesian hierarchical model (BHM) provides a versatile framework to aggregate findings from comparable studies and disentangle genuine variation across studies, resulting from cross-study differences in the respective empirical contexts, from sampling variation in the study-level estimates.

Consider S studies, with associated estimates for the parameter of interest $\hat{\beta}_s$, $s = 1, ..., S$. $\hat{\beta}_s$ may denote for example the estimated degree of overconfidence for either gender, or the gap in overconfidence between genders. The difference between each study-level estimate $\hat{\beta}_s$ and the population mean β can be decomposed into two components. The first component, $\hat{\beta}_s - \beta_s$, represents the difference between study-specific estimates and the respective true values, and reflects sampling (i.e. idiosyncratic) variation, as well as potential biases. The second component, $\beta_s - \beta$, represents the difference between study-specific values and the population value, stemming from systematic differences in the subject population, treatment, or outcomes studied, among other factors.

The above decomposition has two extreme cases. At one extreme, each study identifies a common population effect ($\beta_s = \beta$), and variation across studies is purely idiosyncratic. This is known as the full-pooling (or fixed-effect) model, and has form

$$\hat{\beta}_s \sim N(\beta, \sigma^2), \quad s = 1, ..., S.$$

The estimate of the population mean β is given by the precision-weighted average of the study-level effects:

$$\hat{\beta}^{Pool} = \frac{\sum \hat{\beta}_s / \hat{\sigma}_s^2}{\sum 1/\hat{\sigma}_s^2},$$

where $\hat{\sigma}_s^2$ denotes the variance of each study-level estimate.

Alternatively, in the random-effects model, each study-level estimate $\hat{\beta}_s$ identifies its own study-specific effect β_s, and the study-specific effects are in turn distributed around the population mean, β:

$$\begin{aligned} \hat{\beta}_s &\sim N(\beta_s, \sigma_s^2) \\ \beta_s &\sim N(\beta, \sigma^2), \quad s = 1, ..., S. \end{aligned} \tag{III.1}$$

The estimate of the population parameter β is again a weighted average of the study estimates, in which weights now factor in both the individual study variances $\hat{\sigma}_s^2$ as well as the between-study variance $\hat{\sigma}^2$:

$$\hat{\beta}^{RE} = \frac{\sum \hat{\beta}_s / (\hat{\sigma}_s^2 + \hat{\sigma}^2)}{\sum 1 / (\hat{\sigma}_s^2 + \hat{\sigma}^2)}.$$

In the estimation of the population effect, the random-effects model reduces the precision on all estimates, and relatively more so for more precisely estimated parameters.

The BHM lies between the two extremes. Its formulation resembles that of the random-effects model in (III.1), but – unlike the random-effects model – it treats the "hyperparameters" β and σ^2 as random variables to be estimated:

$$\hat{\beta}_s \sim N(\beta_s, \sigma_s^2) \tag{III.2}$$

$$\beta_s \sim N(\beta, \sigma^2), \quad s = 1, ..., S. \tag{III.3}$$

$$\beta \sim N(-, -) \tag{III.4}$$

$$\sigma^2 \sim N(-, -), \tag{III.5}$$

where $[-, -]$ indicates a prior distribution that needs to be specified. The clear advantage of the BHM is that estimation of σ^2 effectively allows for varying degrees of pooling, where $\sigma^2 = 0$ corresponds to full pooling and $\sigma^2 \to \infty$ corresponds to no pooling.

In the BHM (III.2)-(III.5) we are making a few assumptions. First, condition (III.2) as-

sumes normality of the study effects $\hat{\beta}_s$, which follows from the assumption of internal validity of study-level estimates and the fact that the respective sample sizes are sufficiently large that the central limit theorem can be invoked. Condition (III.3) assumes that the study-level effects are distributed normally around the population mean β. While there is no obvious justification for this assumption, McCulloch and Neuhaus (2011) provide reasonable conditions under which inference on β and σ^2 under the normality assumption is reliable even when the underlying distribution is not normal.

One key assumption in the model above is exchangeability, imposing that the joint distribution of $(\beta_1, ..., \beta_S)$ is invariant to permutations of the indexes $1, ..., S$, allowing us to write the joint distribution of the β_s's as i.i.d. The interpretation of the exchangeability assumption is that studies should be indistinguishable from each other, except for the estimate they provide, such that, for example, there is no reason ex-ante to believe that the estimate from study 1 should be closer to the estimate from study 2 than to the estimate of study 3. This assumption is likely to be violated whenever there are study characteristics that would naturally make some studies more similar to one another than to other studies in the sample. This is clearly the case when multiple estimates are provided within the same paper and are plausibly subject to experimenter effects (Rosenthal 1976).

To address this potential violation of the exchangeability assumption, we introduce an additional layer in the hierarchical model. Our estimation procedure has two steps. In the first step, for each multi-study paper, we estimate a BHM to aggregate information from multiple estimates $k = 1, ..., K$ within each paper s:

$$\begin{aligned} \hat{\beta}_{ks} &\sim N(\beta_{ks}, \sigma^2_{ks}) \\ \beta_{ks} &\sim N(\beta_s, \sigma^2_s), \quad k = 1, ..., K. \end{aligned}$$

In the second step, we use posterior means $\hat{\beta}_s$ and $\hat{\sigma}^2_s$ for multi-study papers, as well as the original estimates from single-study papers, as inputs to the model in (III.2)-(III.5). In the resulting two-step framework, the assumption of exchangeability is imposed within each step.

Finally, as in all Bayesian models, we need to specify a prior distribution on the hyperparameters. In the context of the two-step model, we specify prior distributions for $\hat{\beta}_s$ and $\hat{\sigma}_s^2$ in the first step and for β and σ^2 in the second step. We assume the following prior distributions:

$$\begin{aligned} \beta &\sim N(0,1) \\ \sigma &\sim N(0,1) \\ \beta_s &\sim N(0,0.2^2) \\ \sigma_s &\sim N(0,0.2^2) \end{aligned} \tag{III.6}$$

In all four cases we choose weakly informative priors, which means that we prefer for our posterior distribution (and hence inference), to be driven by information from the data rather than any prior beliefs on overconfidence. The posterior distribution, a function of the prior and the likelihood, is a probability distribution on our parameters of interest, β and σ. In particular, for our second-stage priors we assume that priors for β and σ^2 are normally distributed with zero mean and unit variance. This reflects the fact that absent seeing the data, (1) We have no reason to expect men or women to be fully realistic, or overestimate or underestimate their ability relative to their performance by greater than 1 standard deviation; and (2) We have no reason to expect men to be more overconfident than women, or to be more overconfident than women by greater than 1 standard deviation, or vice versa. The assumption of a zero mean is also consistent with a standard frequentist approach to hypothesis testing, in which the null hypothesis is zero.

Similarly, we assume that our first-stage priors, β_s and σ_s are normally distributed with zero-mean and standard deviation of 0.2. Given that we have a smaller number of experiments within each study,[8] we need more precise priors to regularise the estimates and to prevent over-fitting. This choice may also be justified by noting that there should be smaller heterogeneity in experiments within the same papers, than across papers. The prior standard deviation of 0.2 is similar to the mean standard deviation in the levels of

[8] Across all studies in our analysis, we have a range of 1 to 6 experiments per study

estimated overconfidence within papers (0.19 for women; 0.25 for men), and over twice as large as the mean standard deviation in the corresponding gender gap within multi-study papers (0.097). In the appendix we show that our results are invariant to different choices of scale and location parameters, and functional forms. We also show that our results remain robust to fitting a standard one stage-specification, as in Rubin (1981), in which we treat each experimental observation in the sample, $\hat{\beta}_{ks}$, as exchangeable.

The posterior distribution of the model is proportional to the likelihood and the prior distributions specified above. While we cannot solve for a closed form solution of the posterior distribution, in practice, we characterise the posterior distribution via simulation using Hamiltonian Monte Carlo (HMC), a subset of Markov Chain Monte Carlo (MCMC). HMC methods use derivatives of the density function to construct Markov transitions that sample from the posterior distribution. It does so by introducing auxiliary momentum variables and sampling from a joint density that depends on the auxiliary and posterior distributions. HMCs are more efficient and better suited for estimating hierarchical models than other common MCMC algorithms, including Random Walk Metropolis and Gibbs Sampler (Betancourt and Girolami, 2015; Neal et al., 2011).[9]

Note that the estimated posterior is a joint distribution over not just the population hyperparameters but also each study-level effect. In other words, the best belief about the true effect in a setting is not simply the study-specific estimate. One can in fact improve on the study-specific estimate by factoring in information from $S-1$ comparable studies. This seemingly paradoxical result was first attributed to Charles Stein (Efron and Morris 1977). The intuition behind it is as follows.[10] Consider results from S studies, obtained in S specific settings, $\hat{\beta}_s$, $s = 1, ...S$, and the overall average $\hat{\beta}$. Imagine next to replicate study s in the exact same context. The best prediction for the associated effect is not simply $\hat{\beta}_s$, but indeed it will "shrink" towards the overall average. More generally, all estimates that are above the overall average would be adjusted downward and vice versa. The degree of shrinkage (or pooling) depends on the informative content of each study

[9] We implement this using Stan, a C++ programme that is commonly used for estimating Bayesian models. For each model and metric of interest we use 8 chains and 100,000 iterations per chain.

[10] For the sake of this simple argument, we discuss a one-stage BHM framework.

s about the population of interest. By distinguishing between genuine and sampling variation across studies, the BHM makes this process rigorous and transparent.

To see this more formally, note that in a Normal-Normal hierarchical model specified by equations (III.3)-(III.4) where population parameters β and σ are known, the estimate of the parameter β_s for each study s can be characterized as a shrinkage estimator:

$$\hat{\beta}_s^p = (1 - \lambda_s)\hat{\beta}_s + \lambda_s \hat{\beta}^p$$

where the superscript p denotes posterior estimates and the pooling factor $\lambda_s \in [0, 1]$ captures the degree to which the posterior estimates are shrunk towards the posterior mean.

Following this intuition, Rubin (1981) and Gelman et al. (2013) suggest a pooling metric given by:

$$\hat{\lambda}_s = \frac{\hat{\sigma}_s^{2,p}}{\hat{\sigma}_s^{2,p} + \hat{\sigma}^{2,p}}. \tag{III.7}$$

where $\hat{\sigma_s^{2,p}}$ is the posterior, standard error estimate at the study level and $\hat{\sigma^{2,p}}$ is the corresponding population estimate. In a two-step model, for multi-study papers $\hat{\sigma_s^{2,p}}$ is the posterior estimate from the first stage. For each s, $\lambda_s = 0$ corresponds to full pooling, while $\lambda_s = 1$ corresponds to no pooling. To get an indicator for the degree of pooling at the population level, we estimate $\hat{\lambda}$, which is an arithmetic mean of the pooling metric per study, $\hat{\lambda_s}$.

III.5 Results

Table III.2 summarizes the posterior distribution of the hyperparameters β and σ for male and female overconfidence, and for the gender gap in overconfidence. In all samples, both men and women are found to be overconfident, although only in the full sample does the 95% interval not include zero. In this case, men and women overestimate their performance by about 0.39 and 0.35 standard deviations relative to their ability, respec-

tively. We find little evidence to suggest that men are more overconfident than women. While the estimated gender difference in overconfidence is positive across all three subsamples, its magnitude is small relative to gender-specific means, and in all samples its 95% interval includes zero.

Critically, the results from the BHM suggest that there is a high degree of heterogeneity in the levels and differences in overconfidence across studies. Figure III.5 compares posterior β estimates and their 95% and 90% posterior intervals from each sample to the corresponding full-pooling estimates, which one would obtain under the assumption that each study identifies a common effect. Clearly, the posterior intervals are much wider for the BHM estimates than under the full pooling model, reflecting a high degree of genuine heterogeneity across settings.

This in turn implies that the available body of evidence from the S studies would not be highly-informative about the likely result from the next study, $\hat{\beta}_{S+1}$. Figure III.6 plots the posterior *predictive* distribution for $\hat{\beta}_{S+1}$ for the gender gap in overconfidence in the full sample. Indeed there is 63.8% probability that the next study would find a gender gap in overconfidence ranging (widely) between -0.2 and 0.2.

We next present a more detailed breakdown of results for the full sample,[11] plotting posterior estimates for each study in Figure III.7 for gender-specific overconfidence and Figure III.8 for the gender gap in overconfidence. Compared to the original study estimates, the posterior $\hat{\beta}_s^p$ estimates "shrink" closer to the hyperparameter $\hat{\beta}^p$, plotted at the bottom of each graph. But, as suggested by the comparison between the BHM and full pooling models in Figure III.5, the degree of shrinkage or pooling is quite limited. Table III.3 reports pooling factors, obtained as sample averages of expression (III.7). These imply that only 8% and 6.9% of the variation in estimated overconfidence for men and women, respectively, is explained by sampling variation. For the associated gender gap, the degree of pooling is somewhat higher at 23%. Overall the reported pooling factors imply that the differences in estimates across studies is largely explained by genuine het-

[11] Similar breakdowns for the intensive and extensive margin samples can be found in Figures C6 to C9 in the Appendix.

erogeneity across settings. Thus each additional study on overconfidence tells us little about the overall population mean. In other words, each individual study has limited external validity.

III.6 Explaining the knowledge gap

The BHM estimates indicate that there is no significant gender gap in self-confidence. This however implies a knowledge gap, as experts' opinions are at odds with the BHM analysis of available evidence. The discrepancy is stark. As we have seen, most experts' interpretation of the literature is that there is a positive confidence gap for men whereas the BHM estimates cannot reject a zero gap. On a simple count, Figure III.9 shows that 72% of the findings indicate that both men and women are overconfident, yet only 8% – 26 of 342 respondents – had this interpretation. On the other hand, 77% – 265 respondents – believed that men are overconfident and women underconfident, while only 18% of the findings are in line with this interpretation.

Why are expert economists' beliefs starkly different to the economics literature on confidence? We explore two possible explanations.

First, it is reasonable to hypothesize that highly cited papers play a relatively stronger role in shaping beliefs in the profession. Below we take on board the role of citations by adjusting the estimated precision of each study-level estimate according to its citations. To do so, we estimate a BHM in which we inflate the precision of each study-level estimate by its citations relative to the median number of citations in the sample:

$$\text{Citation Adjusted s.e.} \equiv \tilde{\sigma}_s = \hat{\sigma}_s \times \frac{med(citations_s)}{citations_s}. \tag{III.8}$$

When estimating the posterior mean $\hat{\beta}^p$, this procedure revises upwards the precision of studies with higher than median citations and viceversa.

Table III.4 reports the results obtained, as well as those based on the original standard errors for reference. Comparing estimates in rows 1 and 2, the adjustment does little to

change our main results, and if anything, the posterior mean of 0.080 is slightly smaller than the 0.094 estimate obtained in on the original standard errors. Furthermore, the 95% interval is now wider. The interpretation is that papers finding a larger gender gaps in confidence do not systematically attract more citations.

Can the differences between experts' beliefs and the literature be explained by biased and/or strong prior beliefs on the gender differences in overconfidence? We explore this hypothesis by considering how our estimates change with different assumptions on the moments of the hyper-prior of β, using the standardized mean (1.46) and standard deviation (0.078) of survey responses as a proxy for prior beliefs on the gender differences in overconfidence.

As seen from Table III.5, the gap between the beliefs and the literature can be largely accounted for by an extreme hyper-prior of $\beta \sim N(1.46, 0.078^2)$, wherein the prior belief is not only non-zero but also very strongly held. When we move from our standard model with hyper-priors $\beta \sim N(0,1)$ to one with just a change in the hyper-prior mean ($\beta \sim N(1.46,1)$), the estimated posterior mean and distribution is largely unchanged when compared to our baseline model. However, once we also increase the confidence around the beliefs on the mean, the posterior mean on the average differences in overconfidence almost perfectly coincides with the survey beliefs. This result is intuitive and follows almost mechanically from the set-up of the Bayesian model: given very precise priors, there will be hardly any updating on the posterior mean, regardless of what is found in the literature.[12]

III.7 Discussion

Our analysis yields two main lessons. The first is that the literature in economics provides little support to the hypothesis that differences in self confidence can explain differences in labor market outcomes because, against popular stereotypes, if men are from Mars, so

[12] In Tables C3, C4 and C5 of the appendix, we show that our main findings remain robust to changes on the functional form on priors for β and σ. Tables **??** to C9 show robustness analaysis based on a one-step BHM.

are women. This is important because if men and women do not differ on traits such as confidence, it may be that the barriers/opportunities they face are different and that is what needs to be addressed. However, there is no doubt that in some settings women are less confident than men, but in many others they are not. Indeed, the BHM estimate of the pooling factor is quite low, implying that self-confidence is context specific.

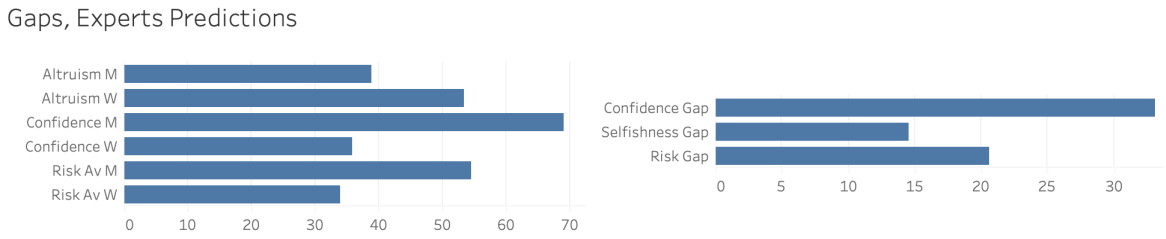
The second, intriguing, finding is that the experts' interpretation of the literature is close to naive pooling and at odds with Bayesian learning. This is especially surprising because for other traits – especially altruism and risk attitudes – the experts' opinions are more in line with BHM estimates. One way to reconcile this is to note that in these domains the pooling factor is high, so that the naive pooling estimate is close to the Bayesian posterior.

This raises the question of how experts learn, because, ultimately, this is what determines the advancement of science.

III.8 Tables and figures

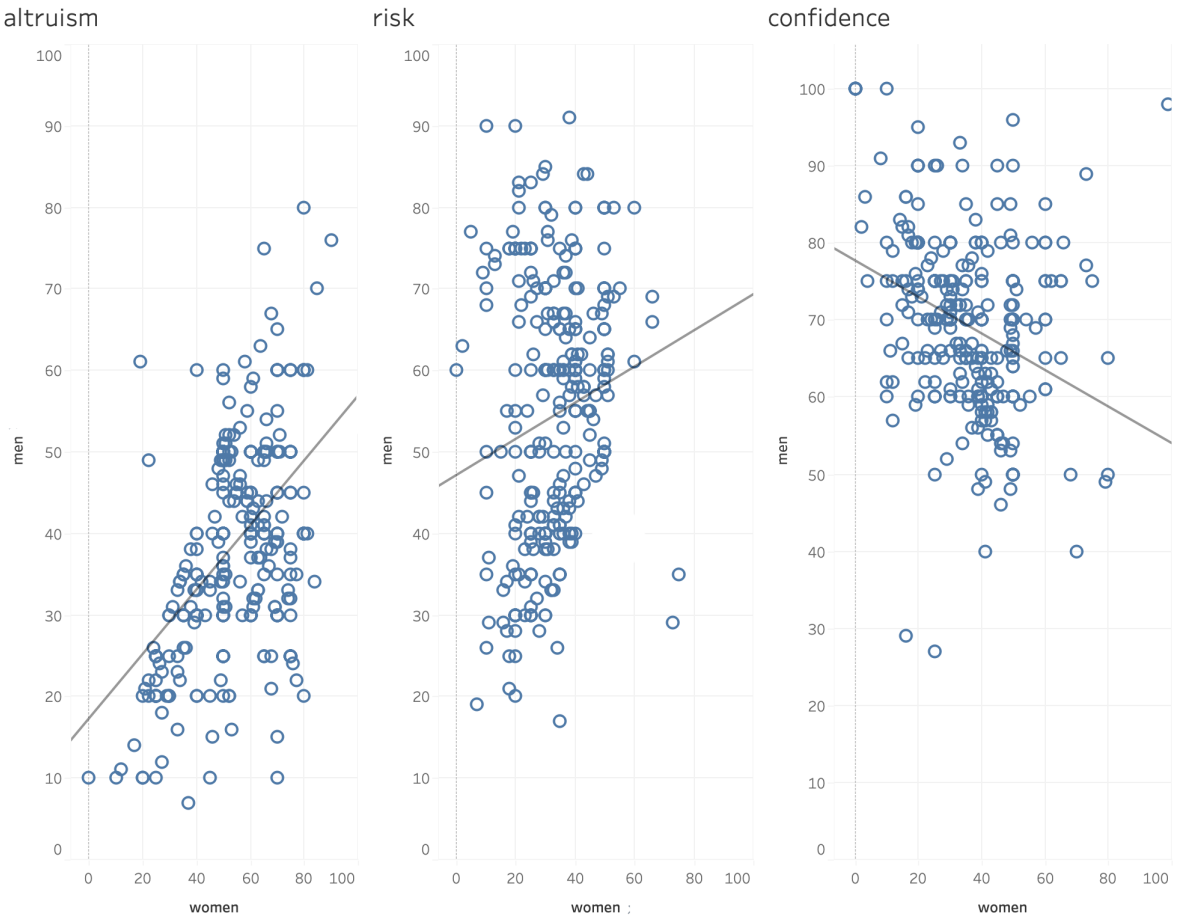
III.8.1 Figures

FIGURE III.1: Experts' answers: Means



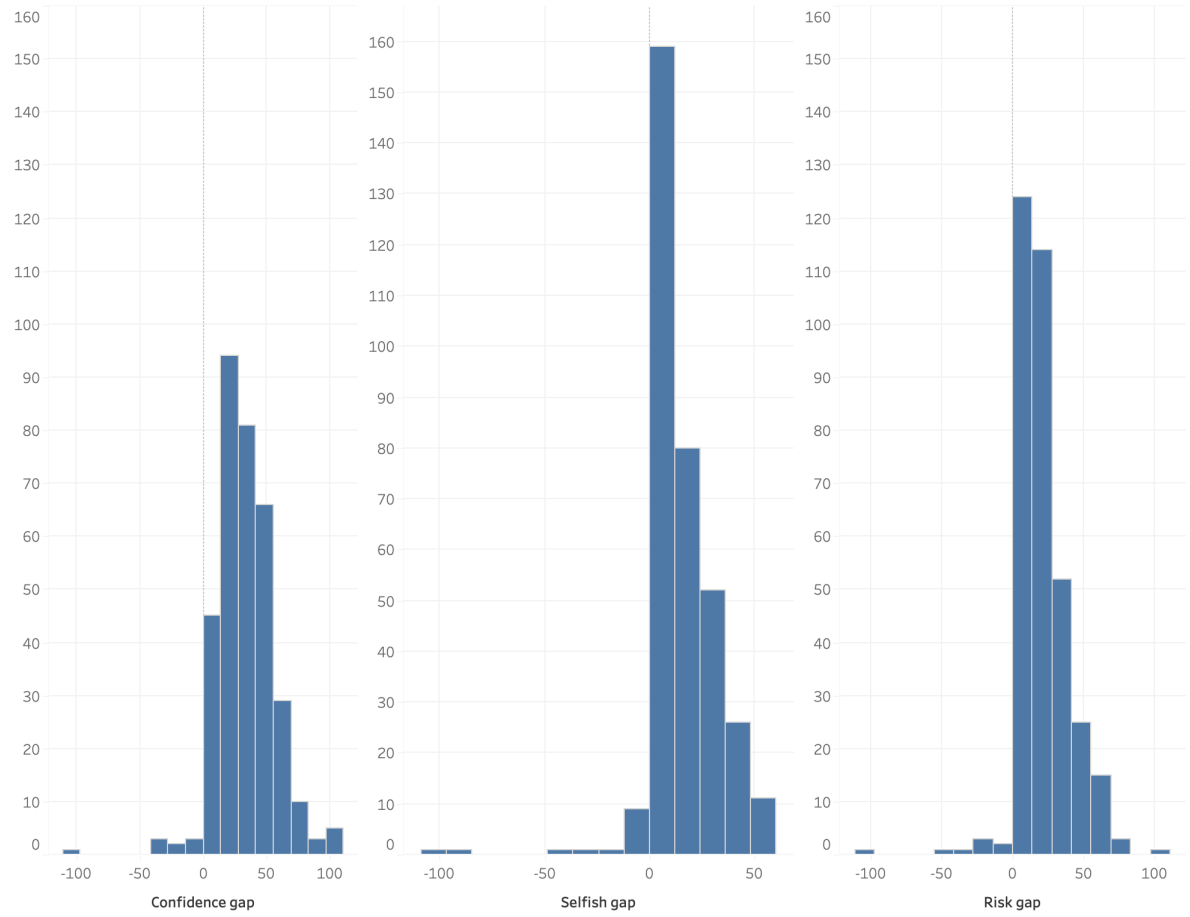
Notes: The panel on the left reports the mean of experts' answers on altruism, self-confidence and risk-aversion. The scale used is described in Figure C1. The panel on the right reports mean gender gaps. N=342.

FIGURE III.2: Experts' answers: Correlations



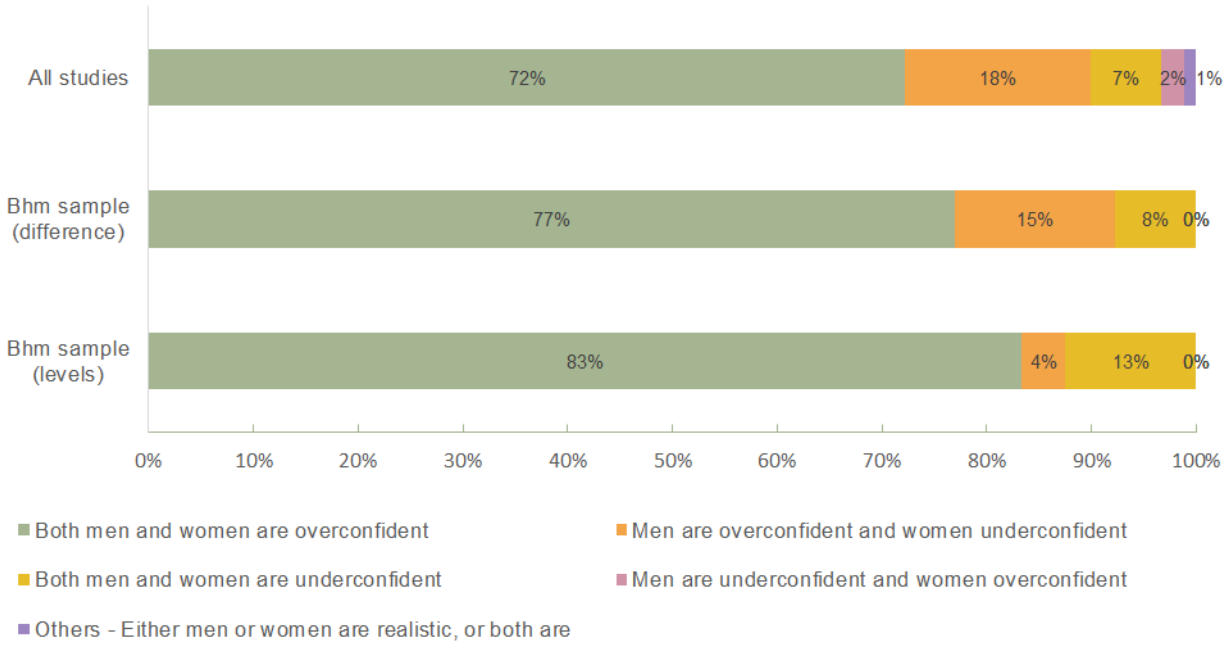
Notes: The graphs plot answers about men against answers about women for each respondent. N=342.

FIGURE III.3: Experts' answers: Distributions



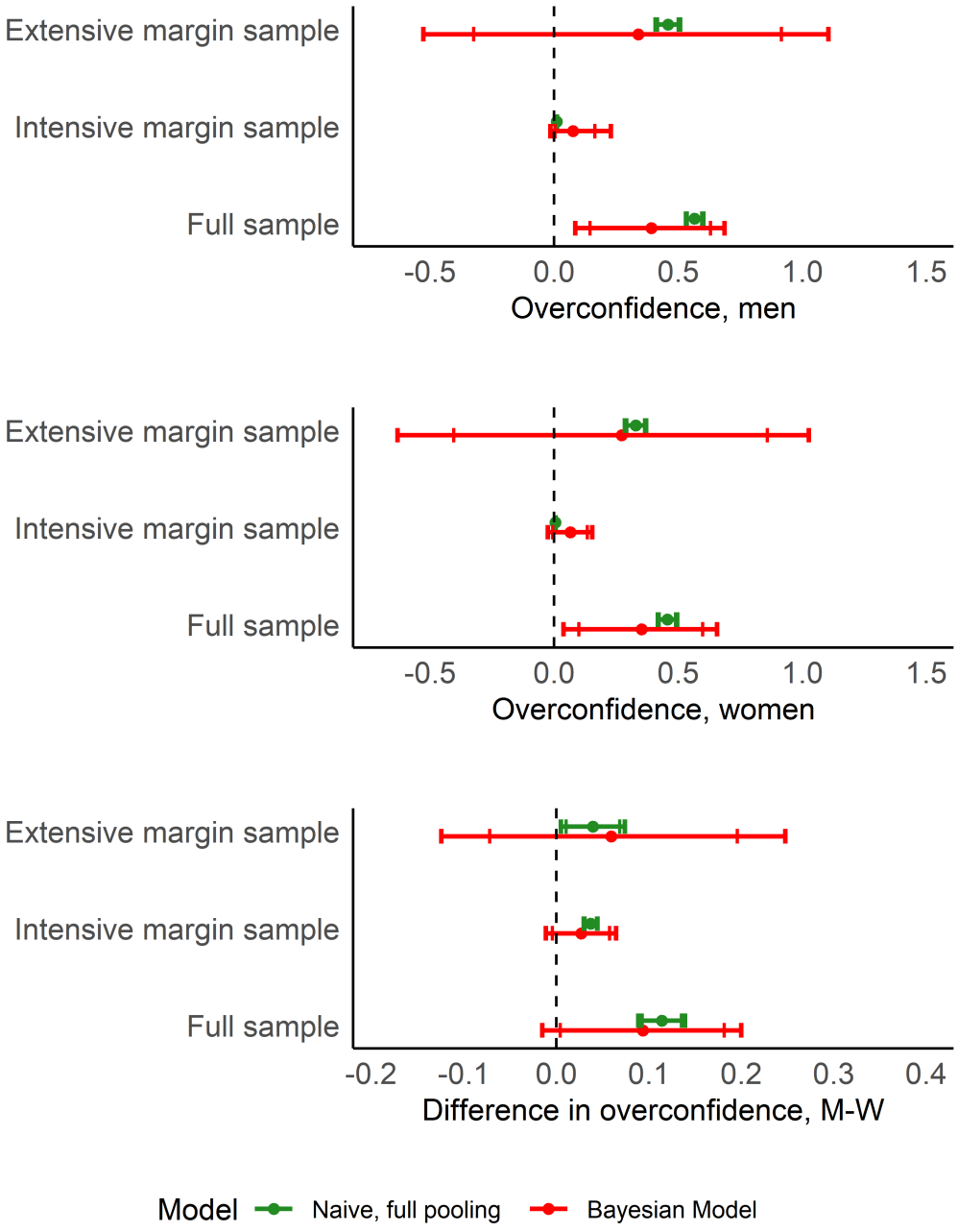
Notes: The histograms display the distribution of answers about gender gaps. N=342.

FIGURE III.4: Distribution of results on self-confidence



Notes: This figure compares the distribution of raw (non-Bayesian aggregated) results across the three samples of the literature on confidence. All studies (N=90) include results with paired observations on confidence for men and women that meet the criteria laid out in the Data section. BHM sample (difference) (N=39) refers to the sub-sample of results that report standard errors for the gender gap in overconfidence. BHM sample (levels) (N=24) refers to the sub-sample of results that report standard errors separately for male and female overconfidence.

FIGURE III.5: Overconfidence of men and women, by sample



Notes: The Figure reports posterior estimates of a two-stage BHM for gender-specific overconfidence and for the corresponding gender difference. The first stage aggregates results at the paper level, and the second stage aggregates paper-level results.

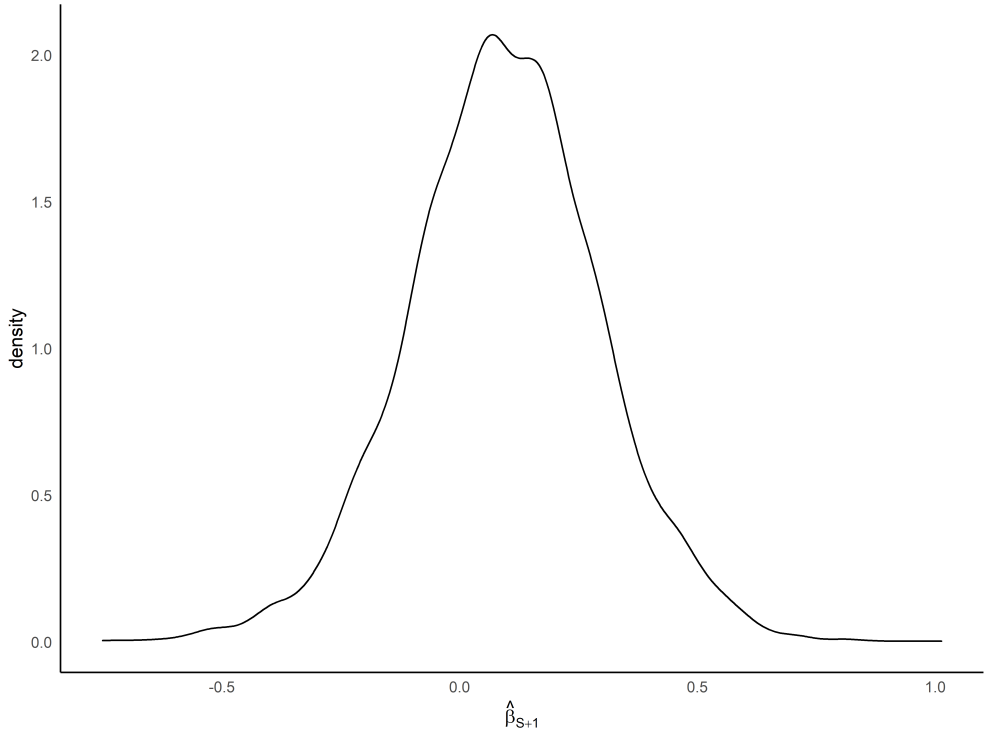
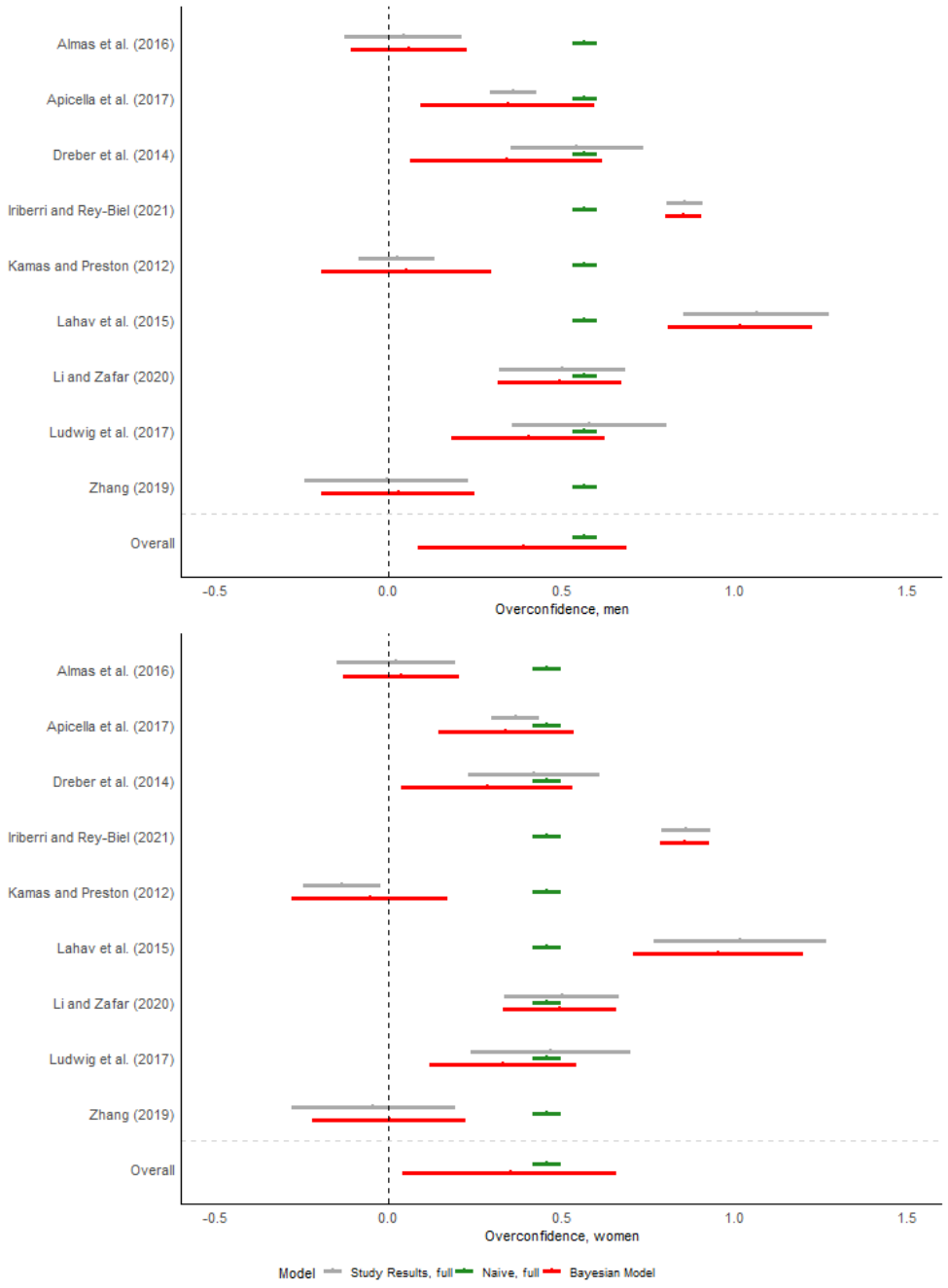


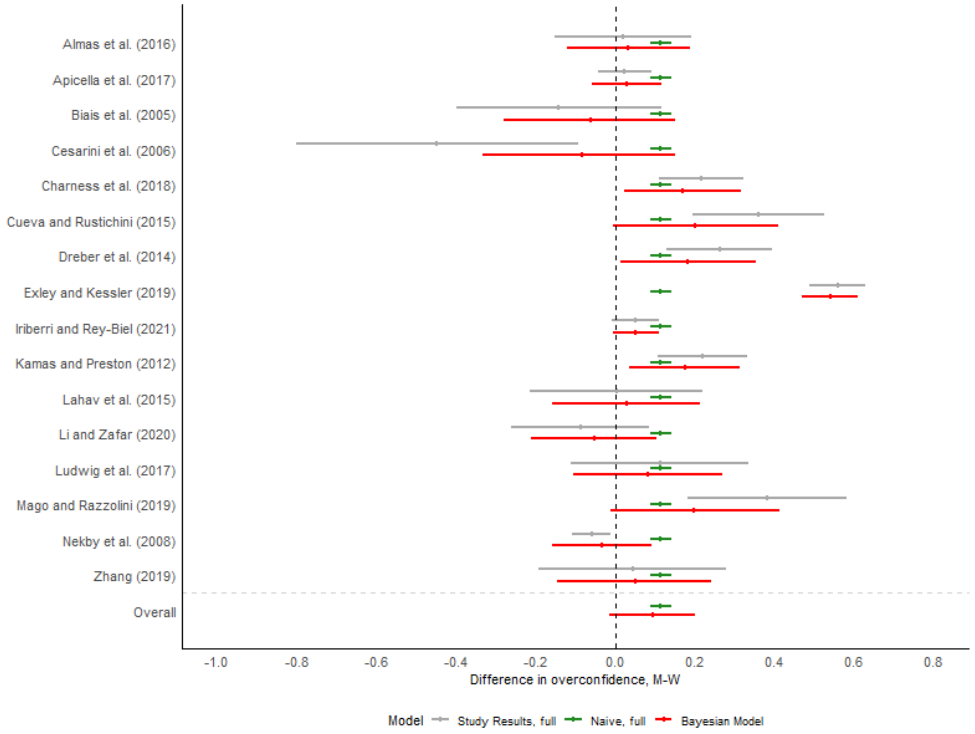
FIGURE III.6: Gender difference in overconfidence, posterior predictive distribution of $\hat{\beta}_{S+1}$

FIGURE III.7: Model comparison - overconfidence, by gender and paper



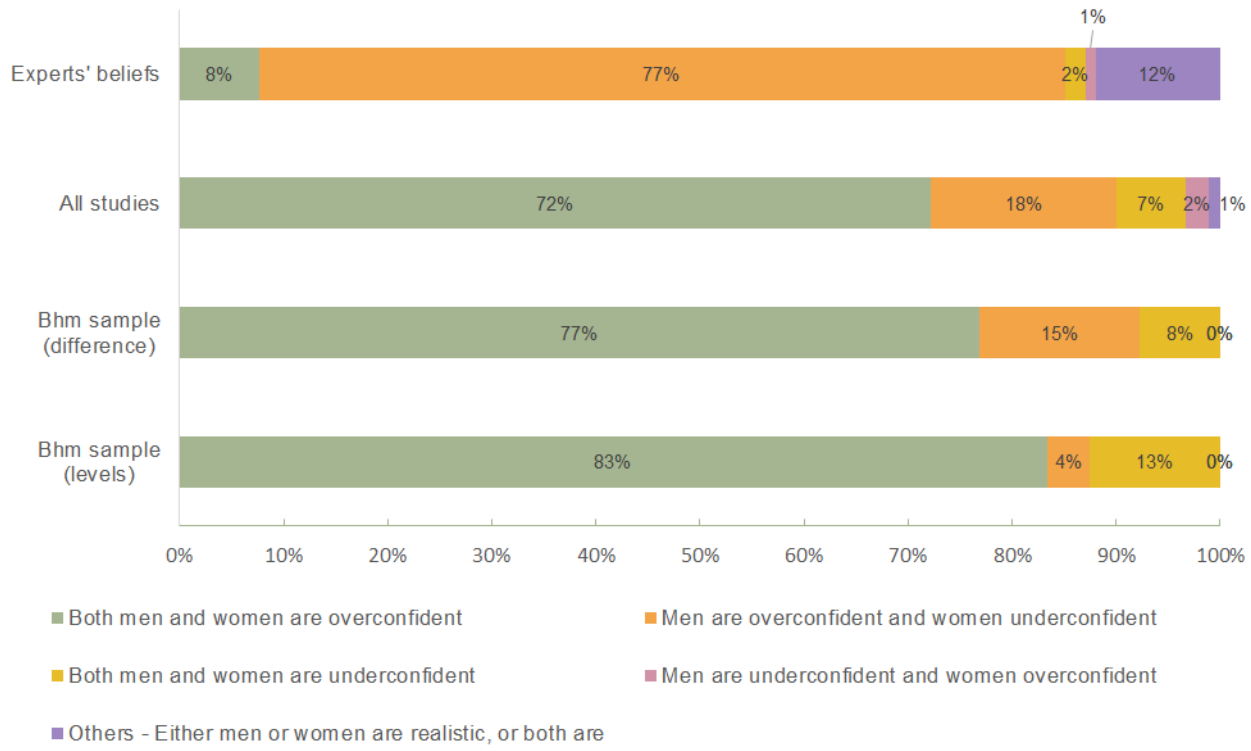
Notes: The Figure reports posterior estimates of a two-stage BHM for gender-specific overconfidence. The first stage aggregates results at the paper level, and the second stage aggregates paper-level results.

FIGURE III.8: Model comparison - Gender differences in overconfidence, by paper



Notes: The Figure reports posterior estimates of a two-stage BHM for gender gaps in overconfidence. The first stage aggregates results at the paper level, and the second stage aggregates paper-level results.

FIGURE III.9: Experts' beliefs vs results on over-confidence



Notes: This figure compares the distribution of expert's beliefs collected from the survey (first bar) to the raw evidence (non-Bayesian aggregated) from the literature (bars 2-4). All studies (N=90) include results with paired observations on confidence for men and women that meet the criteria laid out in the Data section. BHM sample (difference) (N=39) refers to the sub-sample of results that report standard errors for the gender gap in overconfidence. BHM sample (levels) (N=24) refers to the sub-sample of results that report standard errors separately for male and female overconfidence.

III.8.2 Tables

TABLE III.1: Summary evidence on overconfidence

Intensive margin sample			
	N	Mean	St. Dev.
Overconfidence, men	17	0.040	0.073
Overconfidence, women	17	0.027	0.076
Difference (men-women)	26	0.029	0.067
Extensive margin sample			
	N	Mean	St. Dev.
Overconfidence, men	7	0.522	0.186
Overconfidence, women	7	0.462	0.213
Difference (men-women)	13	0.085	0.125
Full sample			
	N	Mean	St. Dev.
Overconfidence, men	24	0.421	0.412
Overconfidence, women	24	0.323	0.372
Difference (men-women)	39	0.115	0.245

Notes: The sample includes 39 studies that report standard errors for the overconfidence metric adopted. The intensive margin subsample includes studies that report the share of men and women who overstate (understate) their performance score (rank). The extensive margin subsample includes studies that report the share of men and women who are overconfident, based on the shares of men and women who overstate (understate) their performance score (rank) or the share of men and women who self-select into a tournament, believing that they will win the tournament, but do not, or the share of men and women who do not self-select into a tournament, believing that they will not win, but would have won based on their performance.

TABLE III.2: Posterior estimates for hyperparameters

	N	J	$\hat{\beta}^p$	$\hat{\sigma}^p$	percentiles				
					2.5	25	50	75	97.5
Intensive margin sample									
Overconfidence, men	17	7	0.076	0.051	-0.015	0.047	0.072	0.098	0.229
Overconfidence, women	17	7	0.065	0.044	-0.026	0.040	0.065	0.090	0.154
Difference (men-women)	26	13	0.027	0.019	-0.011	0.015	0.027	0.039	0.064
Extensive margin sample									
Overconfidence, men	7	2	0.339	0.374	-0.527	0.171	0.363	0.522	1.106
Overconfidence, women	7	2	0.272	0.382	-0.631	0.096	0.301	0.474	1.027
Difference (men-women)	13	4	0.059	0.088	-0.124	0.018	0.058	0.100	0.248
Full sample									
Overconfidence, men	24	9	0.392	0.150	0.086	0.300	0.394	0.485	0.687
Overconfidence, women	24	9	0.352	0.154	0.038	0.259	0.354	0.448	0.657
Difference (men-women)	39	16	0.094	0.054	-0.015	0.059	0.094	0.129	0.200

Notes: The Table reports estimates of a two-stage BHM for gender-specific overconfidence and for the associated gender gap. The first stage aggregates results at the paper level, and the second stage aggregates paper-level results. *N* denotes the number of results included, *J* denotes the number of papers these come from.

TABLE III.3: Pooling factors by metric

Overconfidence of men	$\hat{\lambda} = 0.080$
Overconfidence of women	$\hat{\lambda} = 0.069$
Gender difference in overconfidence	$\hat{\lambda} = 0.23$

*Notes:*The overall pooling factor is obtained as a sample average of expression (III.7).

TABLE III.4: Posterior estimates for hyperparameters based citation-adjusted standard errors

	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	percentiles				
				2.5	25	50	75	97.5
Original s.e.	39	0.094	0.054	-0.016	0.059	0.094	0.129	0.200
Citation adjusted s.e.	39	0.080	0.068	-0.056	0.037	0.081	0.124	0.213

Notes: The Table reports posterior estimates of a two-stage BHM for gender gaps in overconfidence. The first stage aggregates results at the paper level, and the second stage aggregates paper-level results. The standard errors used in estimates of row 2 have been adjusted for cites received, according to expression (III.8).

TABLE III.5: Posterior estimates for hyperparameters for alternative prior distributions

Prior on β	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	percentiles				
				2.5	25	50	75	97.5
$\beta \sim N(0,1)$	39	0.094	0.054	-0.015	0.059	0.094	0.129	0.200
$\beta \sim N(1.46,1)$	39	0.098	0.054	-0.010	0.063	0.098	0.133	0.205
$\beta \sim N(1.46,0.078^2)$	39	1.463	0.006	1.451	1.459	1.463	1.467	1.475

Notes: The table reports posterior estimates of a two-stage BHM for gender gaps in overconfidence. The first stage aggregates results at the paper level, and the second stage aggregates paper-level results.

Bibliography

AGUIAR, F., P. BRAÑAS-GARZA, R. COBO-REYES, N. JIMENEZ, AND L. M. MILLER (2009): "Are women expected to be more generous?" *Experimental Economics*, 12, 93–98.

ALEVY, J. E., F. L. JEFFRIES, AND Y. LU (2014): "Gender-and frame-specific audience effects in dictator games," *Economics Letters*, 122, 50–54.

ALLCOTT, H. (2015): "Site selection bias in program evaluation," *The Quarterly Journal of Economics*, 130, 1117–1165.

ALMÅS, I., A. W. CAPPELEN, K. G. SALVANES, E. Ø. SØRENSEN, AND B. TUNGODDEN (2016): "Willingness to compete: Family matters," *Management Science*, 62, 2149–2162.

ANDREONI, J. AND L. VESTERLUND (2001): "Which is the fair sex? Gender differences in altruism," *The Quarterly Journal of Economics*, 116, 293–312.

ANDREWS, I. AND M. KASY (2019): "Identification of and correction for publication bias," *American Economic Review*, 109, 2766–94.

ANGRIST, J. D. AND J.-S. PISCHKE (2010): "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *Journal of economic perspectives*, 24, 3–30.

APICELLA, C. L., E. E. DEMIRAL, AND J. MOLLERSTROM (2017): "No gender difference in willingness to compete when competing against self," *American Economic Review*, 107, 136–40.

ATTANASIO, O. AND V. LECHENE (2010): "Conditional cash transfers, women and the demand for food," Tech. rep., IFS working papers.

ATTANASIO, O. P. AND V. LECHENE (2014): "Efficient responses to targeted cash transfers," *Journal of political Economy*, 122, 178–222.

AZMAT, G. AND B. PETRONGOLO (2014): "Gender and the labor market: What have we learned from field and lab experiments?" *Labour Economics*, 30, 32–40.

BALAFOUTAS, L. AND M. SUTTER (2010): "Gender, competition and the efficiency of policy interventions," *IZA Discussion Paper*.

BALDIGA, N. R. AND K. B. COFFMAN (2018): "Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women," *Management Science*, 64, 888–901.

BALTRUSCH, M. AND P. C. WICHARDT (2018): "Gender Effects in Dictator Game Giving: Women Favour Female Recipients," .

BANDIERA, O., G. FISCHER, A. PRAT, AND E. YTSMA (2016): "Do women respond less to performance pay? Building evidence from multiple experiments," .

——— (2021): "Do Women Respond Less to Performance Pay? Building Evidence from Multiple Experiments," *American Economic Review: Insights*, Forthcoming.

BANDIERA, O., N. PAREKH, B. PETRONGOLO, AND M. RAO (2022): "Men are from Mars, and Women Too: A Bayesian Meta-analysis of Overconfidence Experiments," *Economica*, 89, S38–S70.

BANERJEE, A., E. DUFLO, N. GOLDBERG, D. KARLAN, R. OSEI, W. PARIENTÉ, J. SHAPIRO, B. THUYSBAERT, AND C. UDRY (2015): "A multifaceted program causes lasting progress for the very poor: Evidence from six countries," *Science*, 348, 1260799.

BANURI, S., S. DERCON, AND V. GAURI (2017): "Biased policy professionals," *The World Bank Economic Review*, 33, 310–327.

BARRERA-OSORIO, F., M. BERTRAND, L. L. LINDEN, AND F. PEREZ-CALLE (2008): "Conditional cash transfers in education design features, peer and sibling effects ev-

idence from a randomized experiment in Colombia," Tech. rep., National Bureau of Economic Research.

BARRIENTOS, A. AND J. M. VILLA (2015): "Evaluating antipoverty transfer programmes in Latin America and sub-Saharan Africa. Better policies? Better politics?" *Journal of globalization and development*, 6, 147–179.

BASTAGLI, F., J. HAGEN-ZANKER, L. HARMAN, V. BARCA, G. STURGE, T. SCHMIDT, AND L. PELLERANO (2016): "Cash transfers: what does the evidence say," *A rigorous review of programme impact and the role of design and implementation features. London: ODI*, 1.

BEAURAIN, G. AND D. MASCLET (2016): "Does affirmative action reduce gender discrimination and enhance efficiency? New experimental evidence," *European Economic Review*, 90, 350–362.

BECKER, A., T. DECKERS, T. DOHMEN, A. FALK, AND F. KOSSE (2011): "The relationship between economic preferences and psychological personality measures," .

BEN-NER, A., F. KONG, AND L. PUTTERMAN (2004): "Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving," *Journal of Economic Psychology*, 25, 581–589.

BEN-NER, A., J. A. LIST, L. PUTTERMAN, AND A. SAMEK (2017): "Learned generosity? An artefactual field experiment with parents and their children," *Journal of Economic Behavior & Organization*, 143, 28–44.

BENJAMIN, D. J., J. J. CHOI, AND A. J. STRICKLAND (2010): "Social identity and preferences," *American Economic Review*, 100, 1913–28.

BENZ, M. AND S. MEIER (2008): "Do people behave in experiments as in the field?—evidence from donations," *Experimental economics*, 11, 268–281.

BERGE, L. I. O., K. BJORVATN, S. GALLE, E. MIGUEL, D. N. POSNER, B. TUNGODDEN,

AND K. ZHANG (2015): "How strong are ethnic preferences?" Tech. rep., National Bureau of Economic Research.

BERTRAND, M. (2011a): "Chapter 17 - New Perspectives on Gender," Elsevier, vol. 4 of *Handbook of Labor Economics*, 1543–1590.

——— (2011b): "New perspectives on gender," in *Handbook of labor economics*, Elsevier, vol. 4, 1543–1590.

——— (2018): "Coase Lecture – The Glass Ceiling," *Economica*, 85, 205–231.

——— (2020): "Gender in the Twenty-First Century," *AEA Papers and Proceedings*, 110, 1–24.

BETANCOURT, M. AND M. GIROLAMI (2015): "Hamiltonian Monte Carlo for hierarchical models," *Current trends in Bayesian methodology with applications*, 79, 2–4.

BEZU, S. AND S. T. HOLDEN (2015): "Generosity and sharing among villagers: Do women give more?" *Journal of Behavioral and Experimental Economics*, 57, 103–111.

BIAIS, B., D. HILTON, K. MAZURIER, AND S. POUGET (2005): "Judgemental overconfidence, self-monitoring, and trading performance in an experimental financial market," *The Review of Economic Studies*, 72, 287–312.

BOLTON, G. E. AND E. KATOK (1995): "An experimental test for gender differences in beneficent behavior," *Economics Letters*, 48, 287–292.

BONARGENT, A. (2024): "Can Research with Policymakers Change the World?" Mimeo.

BORDALO, P., K. COFFMAN, N. GENNAIOLI, AND A. SHLEIFER (2019): "Beliefs about Gender," *American Economic Review*, 109, 739–73.

BOSCHINI, A., A. DREBER, E. VON ESSEN, A. MUREN, AND E. RANEHILL (2018): "Gender and altruism in a random sample," *Journal of behavioral and experimental economics*, 77, 72–77.

BOSCHINI, A., A. MUREN, AND M. PERSSON (2012): "Constructing gender differences in the economics lab," *Journal of Economic Behavior & Organization*, 84, 741–752.

BRANDSTATTER, H. AND W. GUTH (2002): "Personality in dictator and ultimatum games." *Central European Journal of Operations Research*, 10.

BROCK, J. M., A. LANGE, AND E. Y. OZBAY (2013): "Dictating the risk: Experimental evidence on giving in risky environments," *American Economic Review*, 103, 415–37.

BRODEUR, A., M. LÉ, M. SANGNIER, AND Y. ZYLBERBERG (2016): "Star wars: The empirics strike back," *American Economic Journal: Applied Economics*, 8, 1–32.

BUERA, F. J., A. MONGE-NARANJO, AND G. E. PRIMICERI (2011): "Learning the wealth of nations," *Econometrica*, 79, 1–45.

BURKE, M., S. M. HSIANG, AND E. MIGUEL (2015): "Climate and conflict," *Annu. Rev. Econ.*, 7, 577–617.

BUSER, T., M. NIEDERLE, AND H. OOSTERBEEK (2014): "Gender, competitiveness, and career choices," *The Quarterly Journal of Economics*, 129, 1409–1447.

BUSER, T. AND H. YUAN (2019): "Do women give up competing more easily? Evidence from the lab and the Dutch math olympiad," *American Economic Journal: Applied Economics*, 11, 225–52.

CADSBY, C. B., M. SERVÁTKA, AND F. SONG (2010): "Gender and generosity: does degree of anonymity or group gender composition matter?" *Experimental economics*, 13, 299–308.

CAMERER, C. F. AND E. FEHR (2004): "Measuring social norms and preferences using experimental games: A guide for social scientists," *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, 97, 55–95.

CASON, T. N. AND V.-L. MUI (1997): "A laboratory study of group polarisation in the team dictator game," *The Economic Journal*, 107, 1465–1483.

CASTILLO, M. E. AND P. J. CROSS (2008): "Of mice and men: Within gender variation in strategic behavior," *Games and Economic Behavior*, 64, 421–432.

CAVALLO, A., G. CRUCES, AND R. PEREZ-TRUGLIA (2017): "Inflation expectations, learning, and supermarket prices: Evidence from survey experiments," *American Economic Journal: Macroeconomics*, 9, 1–35.

CECCHINI, S. AND B. ATUESTA (2017): "Conditional Cash Transfer Programmes in Latin America and the Caribbean: Coverage and Investment Trends," Available at SSRN: https://ssrn.com/abstract=3037640 or http://dx.doi.org/10.2139/ssrn.3037640.

CESARINI, D., Ö. SANDEWALL, AND M. JOHANNESSON (2006): "Confidence interval estimation tasks and the economics of overconfidence," *Journal of Economic Behavior & Organization*, 61, 453–470.

CHARNESS, G., A. RUSTICHINI, AND J. VAN DE VEN (2018): "Self-confidence and strategic behavior," *Experimental Economics*, 21, 72–98.

CHAUDHRY, T. T. AND M. SALEEM (2011): "Norms of Cooperation, Trust, Altruism, and Fairness: Evidence from Lab Experiments on Pakistani Students." *Lahore Journal of Economics*, 16.

COFFMAN, K. B. (2014): "Evidence on self-stereotyping and the contribution of ideas," *The Quarterly Journal of Economics*, 129, 1625–1660.

COHEN, J. (2013): *Statistical power analysis for the behavioral sciences*, Academic press.

CROSON, R. AND U. GNEEZY (2009a): "Gender differences in preferences," *Journal of Economic literature*, 47, 448–74.

——— (2009b): "Gender Differences in Preferences," *Journal of Economic Literature*, 47, 448–74.

CUEVA, C. AND A. RUSTICHINI (2015): "Is financial instability male-driven? Gender and cognitive skills in experimental asset markets," *Journal of Economic Behavior & Organization*, 119, 330–344.

DASGUPTA, U. (2011): "Do procedures matter in fairness allocations? Experimental evidence in mixed gender pairings," *Economics Bulletin*, 31, 820–829.

DELLAVIGNA, S., W. KIM, AND E. LINOS (2024): "Bottlenecks for Evidence Adoption," *Journal of Political Economy*, 132, 2748–2789.

DREBER, A., E. VON ESSEN, AND E. RANEHILL (2014): "Gender and competition in adolescence: task matters," *Experimental Economics*, 17, 154–172.

DUFLO, E. AND A. BANERJEE (2011): *Poor economics*, vol. 619, PublicAffairs New York, NY, USA.

DUFLO, E. AND M. KREMER (2003): "Use of Randomization in the Evaluation of Development Effectiveness[1]," *Evaluating development effectiveness*, 7, 205.

DUFWENBERG, M. AND A. MUREN (2006): "Generosity, anonymity, gender," *Journal of Economic Behavior & Organization*, 61, 42–49.

DUNNING, T., G. GROSSMAN, M. HUMPHREYS, S. D. HYDE, C. MCINTOSH, AND G. NELLIS (2019): *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*, Cambridge University Press.

DYLONG, P. AND F. KOENINGS (2023): "Framing of economic news and policy support during a pandemic: Evidence from a survey experiment," *European Journal of Political Economy*, 76, 102249.

ECKEL, C. C. AND P. J. GROSSMAN (1998): "Are women less selfish than men?: Evidence from dictator experiments," *The economic journal*, 108, 726–735.

——— (2008): "Differences in the economic decisions of men and women: Experimental evidence," *Handbook of experimental economics results*, 1, 509–519.

EFRON, B. AND C. MORRIS (1977): "Stein's paradox in statistics," *Scientific American*, 236, 119–127.

ENGEL, C. (2011): "Dictator games: A meta study," *Experimental Economics*, 14, 583–610.

EVANS, D. K. AND A. POPOVA (2016): "Cost-effectiveness analysis in development: Accounting for local costs and noisy impacts," *World Development*, 77, 262–276.

EWERS, M. AND F. ZIMMERMANN (2015): "Image and misreporting," *Journal of the European Economic Association*, 13, 363–380.

EXLEY, C. L. AND J. B. KESSLER (2019): "The gender gap in self-promotion," Tech. rep., National Bureau of Economic Research.

FALK, A., A. BECKER, T. J. DOHMEN, D. HUFFMAN, AND U. SUNDE (2016): "The preference survey module: A validated instrument for measuring risk, time, and social preferences," .

FALK, A. AND J. HERMLE (2018): "Relationship of gender differences in preferences to economic development and gender equality," *Science*, 362.

FISZBEIN, A. AND N. R. SCHADY (2009): *Conditional cash transfers: reducing present and future poverty*, World Bank Publications.

FORSYTHE, R., J. L. HOROWITZ, N. E. SAVIN, AND M. SEFTON (1994): "Fairness in simple bargaining experiments," *Games and Economic behavior*, 6, 347–369.

FRANKEL, A. AND M. KASY (2022): "Which findings should be published?" *American Economic Journal: Microeconomics*, 14, 1–38.

FRANZEN, A. AND S. POINTNER (2013): "The external validity of giving in the dictator game," *Experimental Economics*, 16, 155–169.

GALIANI, S. AND P. J. MCEWAN (2013): "The heterogeneous impact of conditional cash transfers," *Journal of Public Economics*, 103, 85–96.

GELMAN, A., J. B. CARLIN, H. S. STERN, D. B. DUNSON, A. VEHTARI, AND D. B. RUBIN (2013): *Bayesian data analysis*, CRC press.

GELMAN, A. AND I. PARDOE (2006): "Bayesian measures of explained variance and pooling in multilevel (hierarchical) models," *Technometrics*, 48, 241–251.

GELMAN, A., D. SIMPSON, AND M. BETANCOURT (2017): "The prior can often only be understood in the context of the likelihood," *Entropy*, 19, 555.

GELMAN, A. ET AL. (2006): "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," *Bayesian analysis*, 1, 515–534.

GERTLER, P. (2004): "Do conditional cash transfers improve child health? Evidence from PROGRESAs control randomized experiment," *American economic review*, 94, 336–341.

GIVEDIRECTLY (2023): "Cash Evidence Explorer," https://www.givedirectly.org/cash-evidence-explorer/, accessed: 2023-06-20.

GOES, I. AND S. B. KAPLAN (2024): "Crude credit: The political economy of natural resource booms and sovereign debt management," *World Development*, 180, 106645.

GONG, B., H. YAN, AND C.-L. YANG (2015): "Gender differences in the dictator experiment: evidence from the matrilineal Mosuo and the patriarchal Yi," *Experimental economics*, 18, 302–313.

GRECH, P. D. AND H. H. NAX (2020): "Rational altruism? On preference estimation and dictator game experiments," *Games and Economic Behavior*, 119, 309–338.

GUMMERUM, M., Y. HANOCH, M. KELLER, K. PARSONS, AND A. HUMMEL (2010): "Preschoolers' allocations in the dictator game: The role of moral emotions," *Journal of Economic Psychology*, 31, 25–34.

GUPTA, N. D., A. POULSEN, AND M. C. VILLEVAL (2005): "Male and female competitive behavior-experimental evidence," *IZA Discussion Paper*.

HADAVAND, A., D. S. HAMERMESH, AND W. W. WILSON (2021): "Publishing Economics: How Slow? Why Slow? Is Slow Productive? Fixing Slow?" Working Paper 29147, National Bureau of Economic Research.

HALVORSEN, T. U. (2015): "Are dictators loss averse?" *Rationality and Society*, 27, 469–491.

HARDIES, K., D. BREESCH, AND J. BRANSON (2013): "Gender differences in overconfidence and risk taking: Do self-selection and socialization matter?" *Economics Letters*, 118, 442–444.

HAUSHOFER, J., P. NIEHAUS, C. PARAMO, E. MIGUEL, AND M. W. WALKER (2022): "Targeting impact versus deprivation," Tech. rep., National Bureau of Economic Research.

HEINZ, M., S. JURANEK, AND H. A. RAU (2012): "Do women behave more reciprocally than men? Gender differences in real effort dictator games," *Journal of Economic Behavior & Organization*, 83, 105–110.

HJORT, J., D. MOREIRA, G. RAO, AND J. F. SANTINI (2021): "How research affects policy: Experimental evidence from 2,150 brazilian municipalities," *American Economic Review*, 111, 1442–1480.

HOUSER, D. AND D. SCHUNK (2009): "Social environments with competitive pressure: Gender effects in the decisions of German schoolchildren," *Journal of Economic Psychology*, 30, 634–641.

HU, M. AND B. LIU (2004): "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177.

HYDE, J. S. (2005): "The gender similarities hypothesis," *American Psychologist*, 60, 581–592.

——— (2014): "Gender Similarities and Differences," *Annual Review of Psychology*, 65, 373–398.

IIDA, Y. (2015): "Task-based income inequalities and redistribution preferences: A comparison of China and Japan," *Journal of Behavioral and Experimental Economics*, 55, 91–102.

INDEPENDENT EVALUATION GROUP (2012): *World Bank Group Impact Evaluations: Relevance and Effectiveness*, Washington, DC: World Bank, license: CC BY 3.0 IGO.

INTERNATIONAL INITIATIVE FOR IMPACT EVALUATION, 3IE (2024): "3ie Development Evidence Portal," Accessed: 2024-11-05.

IRIBERRI, N. AND P. REY-BIEL (2021): "Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment," *European Economic Review*, 131, 103603.

JACKSON, M. O., N. HAGHTALAB, AND A. D. PROCACCIA (2021): "Belief Polarization in a Complex World: A Learning Theory Perspective," *PNAS*, 118.

JAKOBSSON, N. (2012): "Gender and confidence: are women underconfident?" *Applied Economics Letters*, 19, 1057–1059.

JOHN, K. AND S. L. THOMSEN (2017): "Gender differences in the development of other-regarding preferences," .

KAHNEMAN, D., J. L. KNETSCH, AND R. H. THALER (1986): "Fairness and the assumptions of economics," *Journal of business*, S285–S300.

KAMAS, L. AND A. PRESTON (2012): "The importance of being confident; gender, career choice, and willingness to compete," *Journal of Economic Behavior & Organization*, 83, 82–97.

——— (2018): "Competing with confidence: The ticket to labor market success for college-educated women," *Journal of Economic Behavior & Organization*, 155, 231–252.

KAPLAN, S. B. (2018): "Fighting past economic wars: Crisis and austerity in Latin America," *Latin American Research Review*, 53, 19–37.

KASY, M. (2021): "Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It," *Journal of Economic Perspectives*, 35, 175–92.

KITAGAWA, T. AND A. TETENOV (2018): "Who should be treated? empirical welfare maximization methods for treatment choice," *Econometrica*, 86, 591–616.

KLEVEN, H., C. LANDAIS, AND J. E. SØGAARD (2019): "Children and Gender Inequality: Evidence from Denmark," *American Economic Journal: Applied Economics*, 11, 181–209.

KLINOWSKI, D. (2018): "Gender differences in giving in the Dictator Game: the role of reluctant altruism," *Journal of the Economic Science Association*, 4, 110–122.

KREMER, M., M. THOMAS, S. GALLANT, AND O. ROSTAPSHOVA (2021): "Is Development Innovation a Good Investment? Evidence on Scaling and Social Returns from USAID's Innovation Fund," Tech. rep., Working Paper.

LAHAV, E., A. NIR, AND E. SINIVER (2015): "Do differing pay schemes help close the gender gap in overconfidence?"," *Economics Bulletin*, 35, 30–36.

LAZEAR, E. P., U. MALMENDIER, AND R. A. WEBER (2012): "Sorting in experiments with application to social preferences," *American Economic Journal: Applied Economics*, 4, 136–63.

LEIBBRANDT, A., P. MAITRA, AND A. NEELIM (2015): "On the redistribution of wealth in a developing country: Experimental evidence on stake and framing effects," *Journal of Economic Behavior & Organization*, 118, 360–371.

LEVINE, R. AND W. SAVEDOFF (2015): "Aid at the frontier: building knowledge collectively," *Journal of Development Effectiveness*, 7, 275–289.

LEVITT, S. D. AND J. A. LIST (2007): "What do laboratory experiments measuring social preferences reveal about the real world?" *Journal of Economic perspectives*, 21, 153–174.

LI, C. H. AND B. ZAFAR (2020): "Ask and you shall receive? gender differences in regrades in college," *IZA Discussion paper*.

LIST, J. A. (2007): "On the interpretation of giving in dictator games," *Journal of Political economy*, 115, 482–493.

LUDWIG, S., G. FELLNER-RÖHLING, AND C. THOMA (2017): "Do women have more shame than men? An experiment on self-assessment and the shame of overestimating oneself," *European Economic Review*, 92, 31–46.

MAGO, S. D. AND L. RAZZOLINI (2019): "Best-of-five contest: An experiment on gender differences," *Journal of Economic Behavior & Organization*, 162, 164–187.

MARLOWE, F. W. (2004): "What explains Hadza food sharing," *Research in economic Anthropology*, 23, 69–88.

MCCULLOCH, C. E. AND J. M. NEUHAUS (2011): "Misspecifying the shape of a random effects distribution: why getting it wrong may not matter," *Statistical science*, 388–402.

MEAGER, R. (2019): "Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments," *American Economic Journal: Applied Economics*, 11, 57–91.

MEHMOOD, S., S. NASEER, AND D. L. CHEN (2021): "Training policymakers in econometrics," Tech. rep., Technical report. Working Paper.

MENG, X.-L. ET AL. (1994): "Posterior predictive *p*-values," *The annals of statistics*, 22, 1142–1160.

NAKAJIMA, N. (2021): "Evidence-based decisions and Education policymakers," .

NEAL, R. M. ET AL. (2011): "MCMC using Hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, 2, 2.

NEKBY, L., P. S. THOURSIE, AND L. VAHTRIK (2008): "Gender and self-selection into a competitive environment: Are women more overconfident than men?" *Economics Letters*, 100, 405–407.

NIEDERLE, M., C. SEGAL, AND L. VESTERLUND (2013): "How costly is diversity? Affirmative action in light of gender differences in competitiveness," *Management Science*, 59, 1–16.

NIEDERLE, M. AND L. VESTERLUND (2007): "Do women shy away from competition? Do men compete too much?" *The Quarterly Journal of Economics*, 122, 1067–1101.

NIEDERLE, M. AND A. H. YESTRUMSKAS (2008): "Gender differences in seeking challenges: The role of institutions," Tech. rep., National Bureau of Economic Research.

PIKULINA, E., L. RENNEBOOG, AND P. N. TOBLER (2017): "Overconfidence and investment: An experimental approach," *Journal of Corporate Finance*, 43, 175–192.

PROEGER, T. AND L. MEUB (2014): "Overconfidence as a social bias: Experimental evidence," *Economics Letters*, 122, 203–207.

RAO, M. (2020): "Gender differences in altruism: A Bayesian hierarchical analysis of dictator games," *Working paper*.

RAWLINGS, L. B. AND G. M. RUBIO (2005): "Evaluating the impact of conditional cash transfer programs," *The World Bank Research Observer*, 20, 29–55.

REUBEN, E., P. REY-BIEL, P. SAPIENZA, AND L. ZINGALES (2012): "The emergence of male leadership in competitive environments," *Journal of Economic Behavior & Organization*, 83, 111–117.

REUBEN, E., M. WISWALL, AND B. ZAFAR (2017): "Preferences and biases in educational choices and labour market expectations: Shrinking the black box of gender," *The Economic Journal*, 127, 2153–2186.

RIGDON, M., K. ISHII, M. WATABE, AND S. KITAYAMA (2009): "Minimal social cues in the dictator game," *Journal of Economic Psychology*, 30, 358–367.

ROSENTHAL, R. (1976): "Experimenter effects in behavioral research," .

——— (1979): "The file drawer problem and tolerance for null results." *Psychological bulletin*, 86, 638.

ROSENZWEIG, M. R. AND C. UDRY (2020): "External validity in a stochastic world: Evidence from low-income countries," *The Review of Economic Studies*, 87, 343–381.

RUBIN, D. B. (1981): "Estimation in parallel randomized experiments," *Journal of Educational Statistics*, 6, 377–401.

SAAD, G. AND T. GILL (2001): "The effects of a recipient's gender in a modified dictator game," *Applied Economics Letters*, 8, 463–466.

SAMAK, A. C. (2013): "Is there a gender gap in preschoolers' competitiveness? An experiment in the US," *Journal of Economic Behavior & Organization*, 92, 22–31.

SCHULTZ, T. P. (2004): "School subsidies for the poor: evaluating the Mexican Progresa poverty program," *Journal of development Economics*, 74, 199–250.

SIMONSOHN, U., L. D. NELSON, AND J. P. SIMMONS (2014): "P-curve: a key to the file-drawer." *Journal of experimental psychology: General*, 143, 534.

SLONIM, R. AND E. GARBARINO (2008): "Increases in trust and altruism from partner selection: Experimental evidence," *Experimental Economics*, 11, 134–153.

SMITH, A. (2015): "On the nature of pessimism in taking and giving games," *Journal of Behavioral and Experimental Economics*, 54, 50–57.

STEELE, C. M. AND J. ARONSON (1995): "Stereotype threat and the intellectual test performance of African Americans." *Journal of personality and social psychology*, 69, 797.

SUTTER, M., D. GLÄTZLE-RÜTZLER, L. BALAFOUTAS, AND S. CZERMAK (2016): "Cancelling out early age gender differences in competition: an analysis of policy interventions," *Experimental Economics*, 19, 412–432.

SUTTER, M. AND M. G. KOCHER (2001): "Tools for evaluating research output: Are citation-based rankings of economics journals stable?" *Evaluation review*, 25, 555–566.

TOMA, M. AND E. BELL (2024): "Understanding and increasing policymakers' sensitivity to program impact," *Journal of Public Economics*, 234, 105096.

UMER, H. (2020): "Revisiting generosity in the dictator game: Experimental evidence from Pakistan," *Journal of Behavioral and Experimental Economics*, 84, 101503.

USAID (2016): "Strengthening Evidence-Based Development: Five Years of Better Evaluation Practice at USAID 2011–2016," .

VAN RIJN, J., B. BARHAM, AND R. SUNDARAM-STUKEL (2017): "An experimental approach to comparing similarity-and guilt-based charitable appeals," *Journal of behavioral and experimental economics*, 68, 25–40.

VAN RIJN, J., E. J. QUIÑONES, AND B. L. BARHAM (2019): "Empathic concern for children and the gender-donations gap," *Journal of Behavioral and Experimental Economics*, 82, 101462.

VISSER, M. S. AND M. R. ROELOFS (2011): "Heterogeneous preferences for altruism: Gender and personality, social status, giving and taking," *Experimental Economics*, 14, 490–506.

VIVALT, E. (2016): "How much can we generalize from impact evaluations?" .

VIVALT, E. AND A. COVILLE (2023): "How Do Policymakers Update Their Beliefs?" 165, 103121.

WANG, S. AND D. Y. YANG (2021): "Policy Experimentations in China: the Political Economy of Policy Learning," .

WOZNIAK, D., W. T. HARBAUGH, AND U. MAYR (2014): "The menstrual cycle and performance feedback alter gender differences in competitive choices," *Journal of Labor Economics*, 32, 161–198.

YEKUTIELI, D. (2012): "Adjusted Bayesian inference for selected parameters," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 515–541.

ZHANG, Y. J. (2019): "Culture, institutions and the gender gap in competitive inclination: Evidence from the communist experiment in China," *The Economic Journal*, 129, 509–552.

Appendices

A Appendix for Chapter 1

A.1 Additional details on data

Further details on search method

To identify relevant studies in my sample, I replicate the search methodology in Bastagli et al. (2016) for an additional 11 countries in Latin America and the Caribbean in English; and further conduct the same analysis for all countries in my sample in Spanish.

My sample covers all studies published papers (working or final) between 2000 and 2015. The countries included are the following: Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Haiti, Honduras, Jamaica, Mexico, Nicaragua, Panama, Paraguay, Peru, Uruguay.

The search methodology is summarised as follows:

TABLE A1: Search method for program evaluations

	Inclusion Criteria
Keywords	"Cash transfer" + outcome + country name in outlined databases
Outcomes	(1) Monetary poverty, (2) Education, (3) Health and nutrition, (4) Savings, investment, and production, (5) Employment, (6) Empowerment
Databases	EconLit, Scopus, CAB Abstracts, CAB Global Health, POPLINE, Global Health, Google Scholar
Grey literature	World Bank, IFPRI, ECLAC, IADB

Construction of other study characteristics

Earliest date of publication: I identify the earliest date of publication for each study, and assume that this is the date at which policymakers are first aware of the research findings. The method is summarised as follows:

1. Look for the exact citation in google scholar, and check for past or later versions of the paper.

2. IDEAS RePec - contains published and working versions of the paper, especially for those that have been published with international research organisations including IZA, IDB, WB, and IFPRI.

3. Google search of author name + keywords + working paper to identify later or earlier versions of the paper that may have a different name

4. Websites of institutions for the authors of the paper to look for working paper versions of the papers.

5. If no earlier versions of published papers available online, take the full paper submission date for the papers published in journals

Government collaborations with study authors: I identify studies that are conducted in collaboration with government using the following method:

1. Check acknowledgements of the paper for relationships between research project and government institutions.

2. The study is classified as being linked to the government if the research project was funded by or done in collaboration with the researcher or related institution

3. If none above fulfilled, I search for evidence of author and government relationships related to the CCT program at the time of the evaluation or in the years preceding the evaluation years

Demanding and evaluating institutions: Similar to government relationships, I identify the demanding and evaluating agent for each of the evaluations, primarily through the acknowledgements in the evaluation. The demanding agent refers to the type of agent that demands the evaluation. The evaluating agent refers to the type of agent that performs the evaluation.

I classify the identity of the institutions into four categories: (1) research institutions and think tanks; (2) independent researchers; (3) governments; and (4) international institutions. Examples of international institutions include: the World Bank, the IADB, Brooks World Poverty Institute, and the Norwegian Agency for Development Cooperation. I also collect information on the relationship between the demanding and evaluating institution. This gives me a measure of if the evaluation was directly funded by the demanding institution.

A study is classified as being an 'independent' evaluation if it is demanded and conducted by an independent researcher that is not working in collaboration with government.

A.2 Additional results

Individual evaluations and probability of scale-up

TABLE A2: Relationship between measures of evaluation outcomes and probability of scale-up, defined as greater than 10% increase in spending

	Dependent variable: $1(ScaleUp > 10\%)$				
	Measure of evaluation outcome				
	Mean t-stat	Max t-stat	Mean effect size	Max effect size	Abstract sentiment
Constant	0.3893	0.3811	0.4270	0.4296	0.3968
	(0.0731)	(0.0900)	(0.0702)	(0.0699)	(0.0728)
TE_{it-1}	0.0088	0.0072	-0.5296	-0.2386*	1.5268
	(0.0127)	(0.0137)	(0.3144)	(0.1149)	(3.2173)
Observations	105	105	105	105	64
R^2	0.00133	0.00163	0.01544	0.01769	0.00646
Adjusted R^2	-0.00836	-0.00806	0.00588	0.00816	-0.00957

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and probability of scale-up, as defined as a spending increase greater than 10%. The evaluation results (TE_{it-1}) in each study are summarised by: (1) the mean t-statistic of headline results; (2) maximum t-statistic of headline result; (3) mean effect size of headline results; (4) maximum effect size of headline results; and (5) the abstract sentiment. Standard errors are clustered at the country level.

TABLE A3: Relationship between measures of evaluation outcomes and probability of scale-up, defined as greater than 20% increase in spending

	Dependent variable: $1(ScaleUp > 20\%)$				
	Measure of evaluation outcome				
	Mean t-stat	Max t-stat	Mean effect size	Max effect size	Abstract sentiment
Constant	0.2684	0.2484	0.3017	0.2981	0.2902
	(0.0848)	(0.0897)	(0.0773)	(0.0695)	(0.0598)
TE	0.0064	0.0105	-0.5016	-0.1764	1.0797
	(0.0228)	(0.0144)	(0.3456)	(0.0972)	(1.7601)
Observations	105	105	105	105	64
R^2	0.00085	0.00425	0.01663	0.01160	0.00373
Adjusted R^2	-0.00885	-0.00541	0.00709	0.00201	-0.01234

Notes: This table shows the linear relationship (OLS) between the treatment effect, TE_{it-1} from study i, first published in year $t-1$, and probability of scale-up, as defined as a spending increase greater than 20%. The evaluation results (TE_{it-1}) in each study are summarised by: (1) the mean t-statistic of headline results; (2) maximum t-statistic of headline result; (3) mean effect size of headline results; (4) maximum effect size of headline results; and (5) the abstract sentiment. Standard errors are clustered at the country level.

Cumulative evidence and spending in 2015

TABLE A4: Relationship between posterior mean of aggregate findings and CCT spending, 2015

	log(CCT spend)	CCT spend as % of social protection	CCT spend as % of GDP
Constant	19.6391	0.1427	0.0034
	(0.5600)	(0.0669)	(0.0006)
Posterior mean	-0.2455	0.0122	-0.0005
	(0.4036)	(0.0568)	(0.0004)
Observations	16	16	16
R^2	0.02047	0.00409	0.09086
Adjusted R^2	-0.04949	-0.06705	0.02592

Notes: Linear relationship between posterior mean of aggregate treatment effects for each country, and measures of CCT spending in 2015. Posterior mean is estimated from the Bayesian hierarchical model, using aggregate evidence on CCTs in each country, between 2000 to 2015.

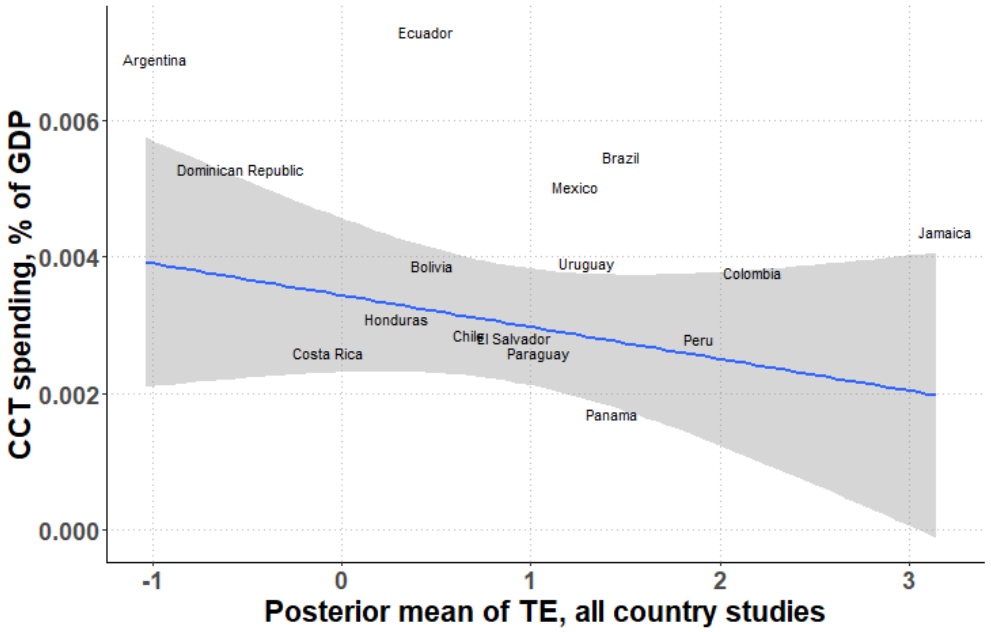


FIGURE A1: Posterior mean of treatment effects and spending, as percentage of GDP

Notes: Linear relationship between posterior mean of aggregate treatment effects for each country, and CCT spending as a percentage of GDP in 2015. Posterior mean is estimated from the Bayesian hierarchical model, using aggregate evidence on CCTs in each country, between 2000 to 2015.

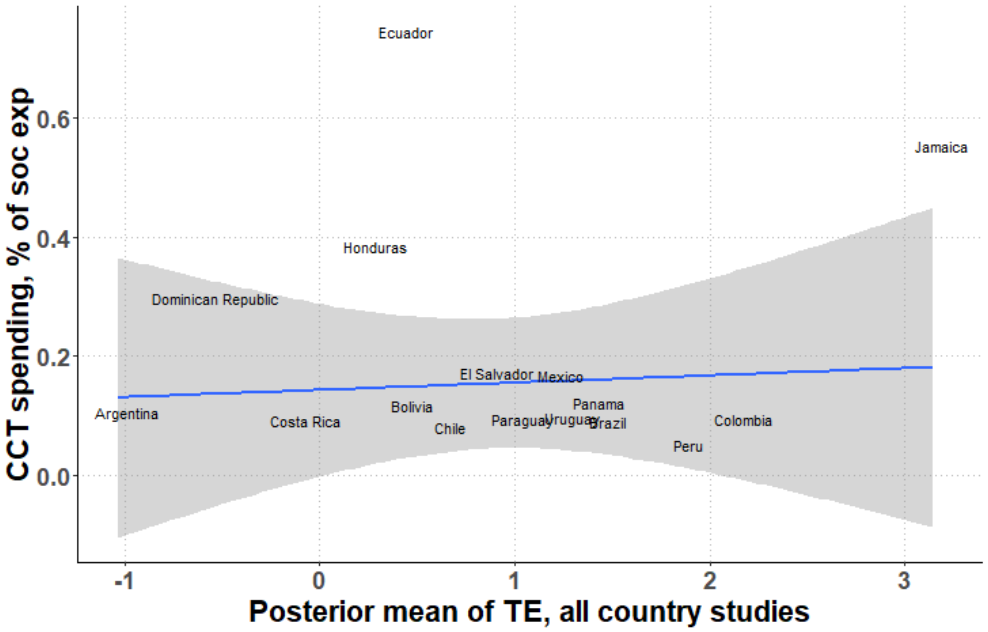


FIGURE A2: Posterior mean of treatment effects and spending, as percentage of social protection

Notes: Linear relationship between posterior mean of aggregate treatment effects for each country, and CCT spending as a percentage of social protection expenditure in 2015. Posterior mean is estimated from the Bayesian hierarchical model, using aggregate evidence on CCTs in each country, between 2000 to 2015.

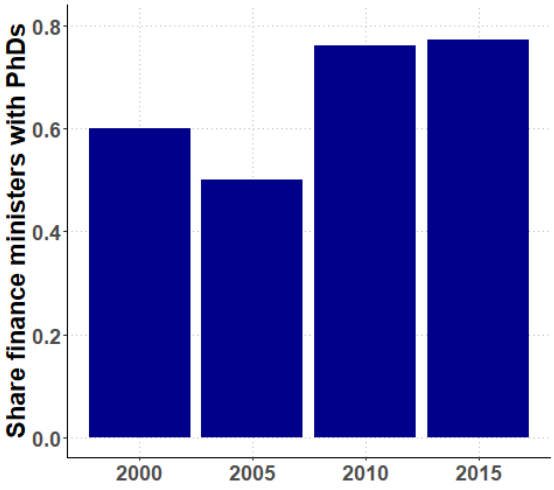


FIGURE A3: Proportion of finance ministers with PhDs in Latin America and the Caribbean, by year

Notes: This figure shows the proportion of finance ministers in LAC countries with PhDs. Estimates using data from the Index of Economic Advisers, (Goes and Kaplan, 2024; Kaplan, 2018).

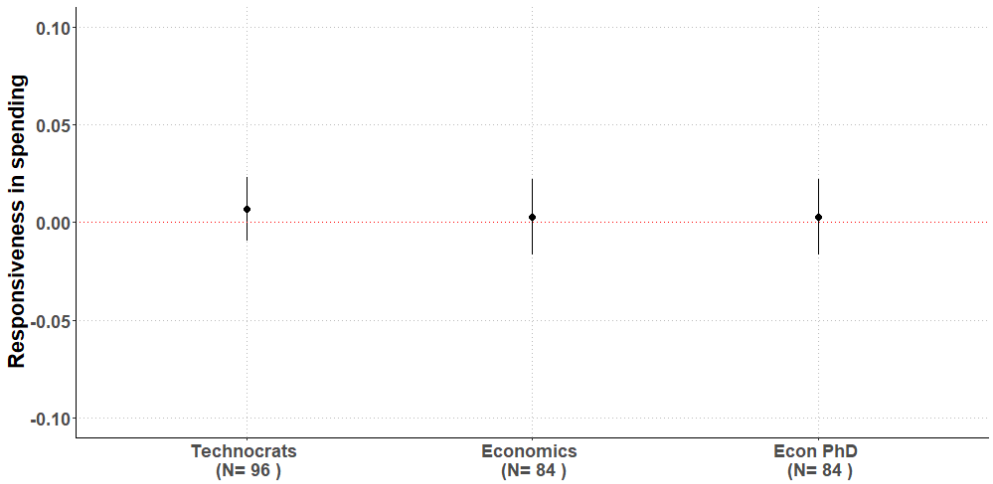


FIGURE A4: Relationship between mean t-stat and spending, by finance minister training

Notes: This figure shows the linear relationship between evaluation outcomes (mean t-statistic) and spending, one year after first publication date, by the training of ministers at first publication date. Technocrats, are those with PhDs; Economics, are those with economics degrees (including graduate and undergraduate studies); and Econ Phds are exclusively economics PhDs.

Robustness of timeliness of evaluation

I consider robustness of the results to assumptions around when policymakers may first become aware of the research results. This may be a concern primarily for studies that are more timely, i.e. released within four years of the effect year.

In the figure below, I consider the relationship between evaluation outcomes and subsequent changes in spending for timely evaluations, where I assume that the first date at which policymakers may be aware of the evidence is the effect year. The results suggest that the positive association between evaluation outcomes and spending for timely evaluations in figure I.19 is not driven by policymakers incorporating findings to their spending decisions prior to the research results being published.

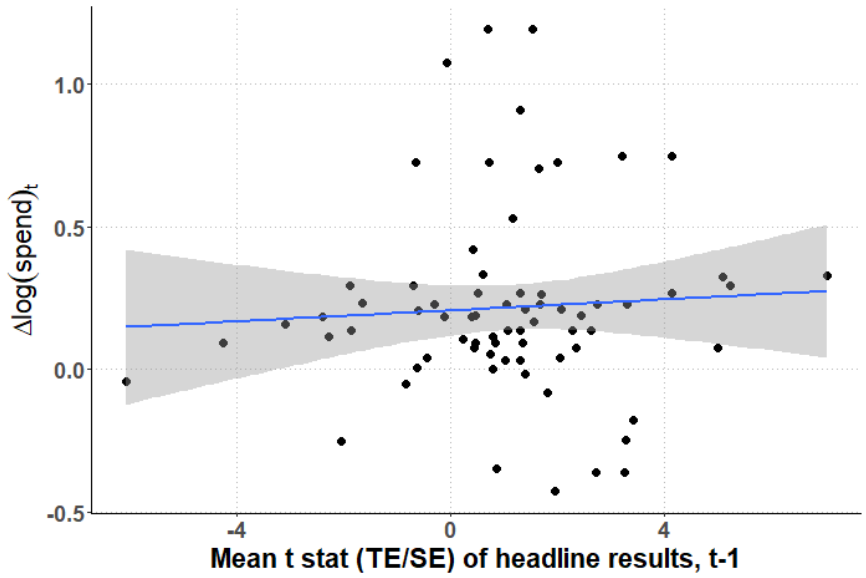


FIGURE A5: Relationship between mean t-stat and subsequent spending, matched by the endline year of the evaluation. Timely evaluations only.

Notes: Linear relationship between mean t-stat and changes in spending, one year after the effect year of evaluation. Consider only timely evaluations, i.e. subset of evaluations released within four years of the effect year.

Lastly, since several of the non-timely studies involve re-analyses of experimental data from previous studies (e.g. PROGRESA), I examine whether the findings on the importance of timeliness are driven by the subset of studies that are re-analysis of existing data.

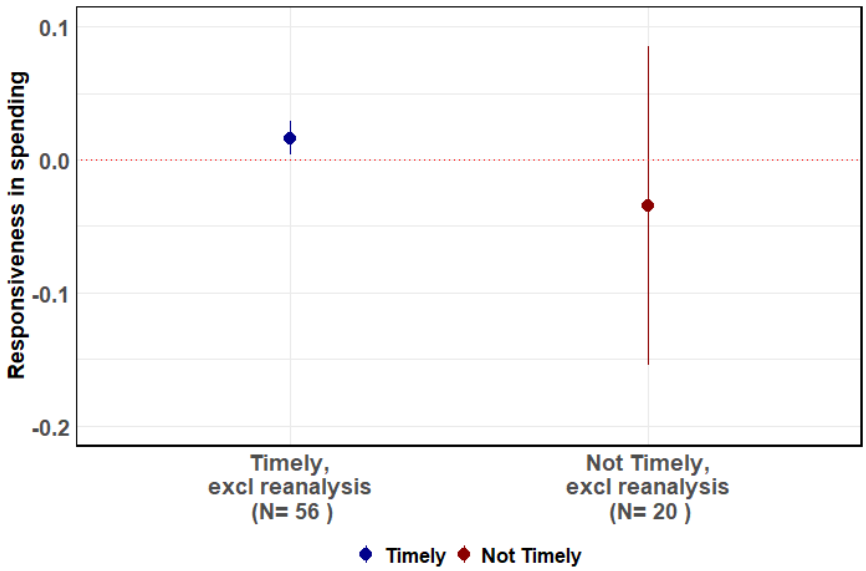


FIGURE A6: Relationship between mean tstat and subsequent spending, excluding studies that use experimental data from prior RCTs

Notes: Linear relationship between program evaluation outcomes and changes in spending, one year after the evaluation is first made available, by timeliness of the study. These results exclude the subset of studies that are re-analyses of experimental data from past studies.

B Appendix for Chapter 2

B.1 Robustness analysis

The validity of posterior inference is critically dependent on the set of assumptions on the probability model, as laid out in Section II.3.1. It is thus necessary to assess the fit and sensitivity of the model to these assumptions. In this section, I conduct a series of posterior predictive checks and explore the sensitivity of the analysis to different distributional assumptions on the priors.

Posterior predictive checks

If the model is suited to the setting, simulations under the posterior predictive distribution should look similar to the distribution of the true data. That is, after estimation, it should seem plausible that the data was generated with the chosen model (Gelman et al., 2013). While the use of posterior predictive checks violates the likelihood principle, in that the data is being used twice (for estimation and for model checking), Meng et al. (1994) and Gelman et al. (2013) argue that, at the very least, we should look for systematic differences between the data and simulations from the posterior predictive distribution to understand the limitations of the model.

In Figure B1, I overlay the cumulative density of the data with that of simulations from the posterior predictive distribution. For simplicity, I suppress the subscripts and let y denote observations from my data, and y^{rep} denote simulations of the data from the posterior predictive distribution. The cumulative density of the simulated data closely resembles that of the true data, suggesting that it is plausible that the data could be generated by the model.

I further construct measures of the fit by considering a series of relevant test statistics for the posterior predictive distribution. For each test quantity, $T(\hat{y})$, I calculate the

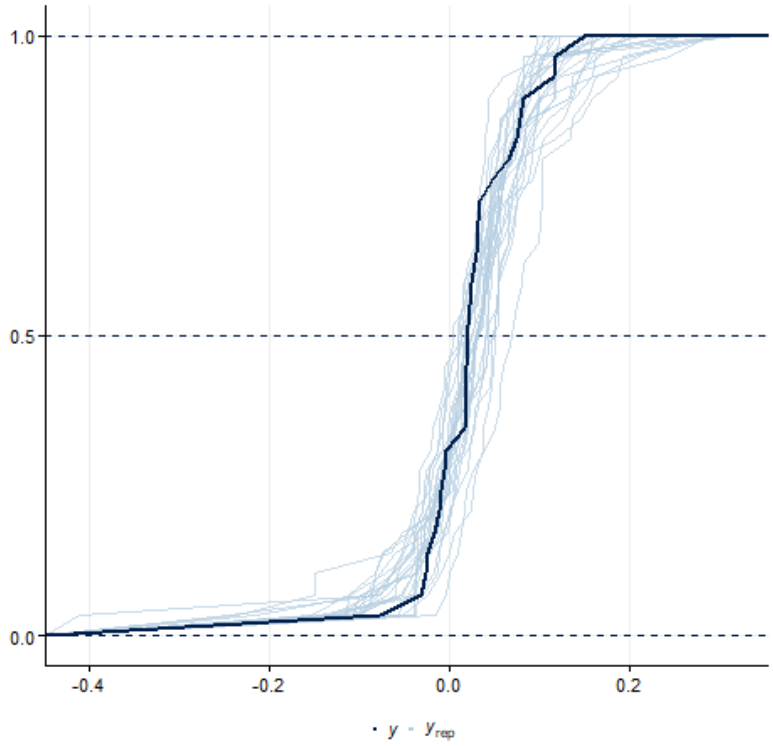


FIGURE B1: Cumulative density of data versus simulated data

Notes: This figure plots the cumulative density of data, y, overlaid with cumulative density of 25 simulations from the posterior predictive distribution, y^{rep} .

corresponding Bayesian p-value, p_b as follows:

$$p_b = Pr(T(\hat{y}^{rep}, \theta) \geq T(\hat{y}, \theta) \mid \hat{y})$$

In practice, the Bayesian p-value is calculated as the proportion of simulations from the posterior predictive distribution, for which the simulated value of the test statistic is greater than the test quantity calculated from the data. The closer is the p-value to 0 or to 1, the less likely it is that the data would be generated under the posterior predictive distribution implied by the model.

In Figure B2 I consider four test-statistics of interest: the maximum, minimum, median, and mean of study effects. I plot the posterior predictive distributions for each of these test statistics, using the value of the test statistic for 1000 simulations of the predictive data. For each of these, the Bayesian p-value is sufficiently far away from 0 and 1, which suggests that the model generates predicted values that are close to the sample data.

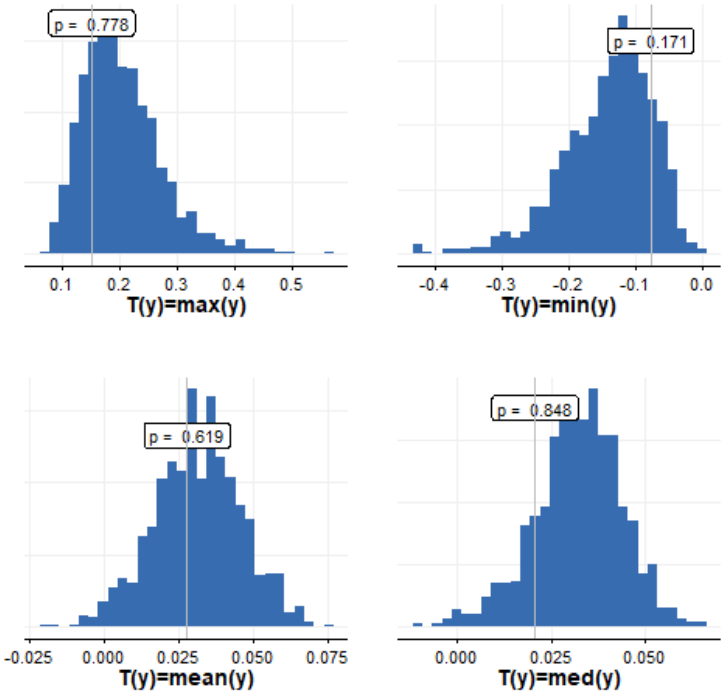


FIGURE B2: Posterior predictive distribution and associated p-value for four test statistics.

Notes: Vertical lines denote the value of the test statistic for the data.

Prior checks

A second concern on inference is the sensitivity of results to the choice of the prior distribution. In Table B1 I explore the sensitivity of my estimates to 12 alternative choices of the prior distribution. For each of these specifications, I center the prior distribution around a zero mean, consistent with the assumption of a null effect unless proven otherwise by the data (as is the approach with hypothesis testing). The posterior mean and 95% interval for μ remain stable for the range of different distributional assumptions.

TABLE B1: Prior checks - estimates of posterior mean

Model Priors	Mean	s.e.	2.5%	50%	97.5%
$\mu \sim$normal(0,1); $\tau \sim$normal(0,1)	0.0324	0.0072	0.0171	0.0328	0.0454
$\mu \sim$cauchy(0,1); $\tau \sim$normal(0,1)	0.0328	0.0069	0.0184	0.0329	0.0458
$\mu \sim$normal(0,10); $\tau \sim$normal(0,1)	0.0324	0.0072	0.0172	0.0327	0.0457
$\mu \sim$cauchy(0,10); $\tau \sim$normal(0,1)	0.0324	0.0072	0.0173	0.0328	0.0455
$\mu \sim$normal(0,1); $\tau \sim$normal(0,10)	0.0326	0.0070	0.0175	0.0328	0.0455
$\mu \sim$cauchy(0,1); $\tau \sim$normal(0,10)	0.0323	0.0072	0.0171	0.0326	0.0453
$\mu \sim$normal(0,10); $\tau \sim$normal(0,10)	0.0320	0.0073	0.0159	0.0324	0.0449
$\mu \sim$cauchy(0,10); $\tau \sim$normal(0,10)	0.0325	0.0070	0.0174	0.0330	0.0456
$\mu \sim$normal(0,1); $\tau \sim$uniform(0,1)	0.0326	0.0067	0.0188	0.0329	0.0450
$\mu \sim$cauchy(0,1); $\tau \sim$uniform(0,1)	0.0323	0.0077	0.0170	0.0329	0.0451
$\mu \sim$normal(0,10); $\tau \sim$uniform(0,10)	0.0327	0.0070	0.0178	0.0329	0.0457
$\mu \sim$cauchy(0,10); $\tau \sim$uniform(0,10)	0.0325	0.0071	0.0172	0.0329	0.0459

Notes: Posterior estimates of gender difference in contributions, across hyperpriors assumptions.

B.2 Papers on dictator games in sample

TABLE B2: Summary of studies and experiment characteristics

Study	Relevant Journal	Gender main topic	Number of observations	Share women	Number of relevant experiments	Source of variation in experiments, if multiple
Aguiar et al. (2009)	Yes	Yes	40	0.50	1	
Alevy et al. (2014)	No	Yes	219	0.50	4	Anonymity of dictator, framing
Andreoni and Vesterlund (2001)	Yes	Yes	1136	0.33	8	Price of giving, size of pie
Baltrusch and Wichardt (2018)	No	Yes	1016	0.27	2	Identity of recipient
Ben-Ner et al. (2017)	Yes	No	293	0.67	2	Anonymity of dictator
Berge et al. (2015)	Yes	No	4048	0.60	5	Anonymity of dictator, timing of game
Bezu and Holden (2015)	Yes	Yes	724	0.50	2	Identity of recipient
Boschini et al. (2012)	Yes	Yes	1086	0.64	12	Priming (gender), price of giving
Boschini et al. (2018)	Yes	Yes	889	0.40	4	Priming (gender)
Brandstatter and Guth (2002)	No	No	51	0.61	1	
Brock et al. (2013)	Yes	No	63	0.46	1	
Cadsby et al. (2010)	Yes	Yes	699	0.49	4	Anonymity of dictator
Cason and Mui (1997)	Yes	No	188	NA	1	
Castillo and Cross (2008)	Yes	Yes	107	0.41	4	Price of giving, size of pie
Chaudhry and Saleem (2011)	No	No	238	0.56	1	
Dasgupta (2011)	No	Yes	80	0.50	1	
Dufwenberg and Muren (2006)	Yes	Yes	352	0.48	2	Anonymity of dictator
Eckel and Grossman (1998)	Yes	Yes	120	0.50	1	
Gong et al. (2015)	Yes	Yes	144	0.50	2	Subject population
Grech and Nax (2020) 2019	Yes	No	4120	0.61	1	
Gummerum et al. (2010)	No	No	77	0.55	1	
Halvorsen (2015)	No	No	177	0.40	4	Framing
Heinz et al. (2012)	Yes	Yes	83	0.55	2	Size of pie
Houser and Schunk (2009)	No	Yes	151	0.47	3	Anonymity of dictator
Iida (2015)	Yes	No	168	0.30	2	Subject population
John and Thomsen (2017)	Yes	Yes	985	0.48	1	
Klinowski (2018)	No	Yes	308	0.50	1	

TABLE B2: Summary of studies and experiment characteristics (continued)

Study	Relevant Journal	Gender main topic	Number of observations	Share women	Number of relevant experiments	Source of variation in experiments, if multiple
Lazear et al. (2012)	Yes	No	83	0.53	1	
Leibbrandt et al. (2015)	No	No	90	0.33	4	Size of pie, framing
Marlowe (2004)	No	No	43	0.49	1	
Rigdon et al. (2009)	No	No	113	0.55	2	
Saad and Gill (2001)	No	Yes	224	0.48	2	Identity of recipient
Slonim and Garbarino (2008)	Yes	No	580	0.52	2	Identity of recipient
Smith (2015)	Yes	No	144	0.21	3	Framing
Umer (2020)	Yes	No	157	0.50	2	Anonymity of dictator
van Rijn et al. (2017)	Yes	No	166	0.72	1	
van Rijn et al. (2019)	Yes	Yes	573	0.54	4	Priming (guilt)
Visser and Roelofs (2011)	Yes	Yes	530	0.65	5	Price of giving, size of pie
Overall (out of 38 studies)	25	20	20,265	0.53	100	

Notes: Relevant Journals are defined as papers published in the top 5 economics journals, the field journals in behavioral and experimental economics, and the main working paper series (NBER, IZA, and CEPR). See Footnote 2 for full list. Gender listed as main topic if a gender related term (e.g. women, men, gender) is used in the title of the paper. Cason and Mui (1997) do not report gender split of participants.

C Appendix for Chapter 3

C.1 Additional statistics: Expert's survey

FIGURE C1: Expert survey questions 1

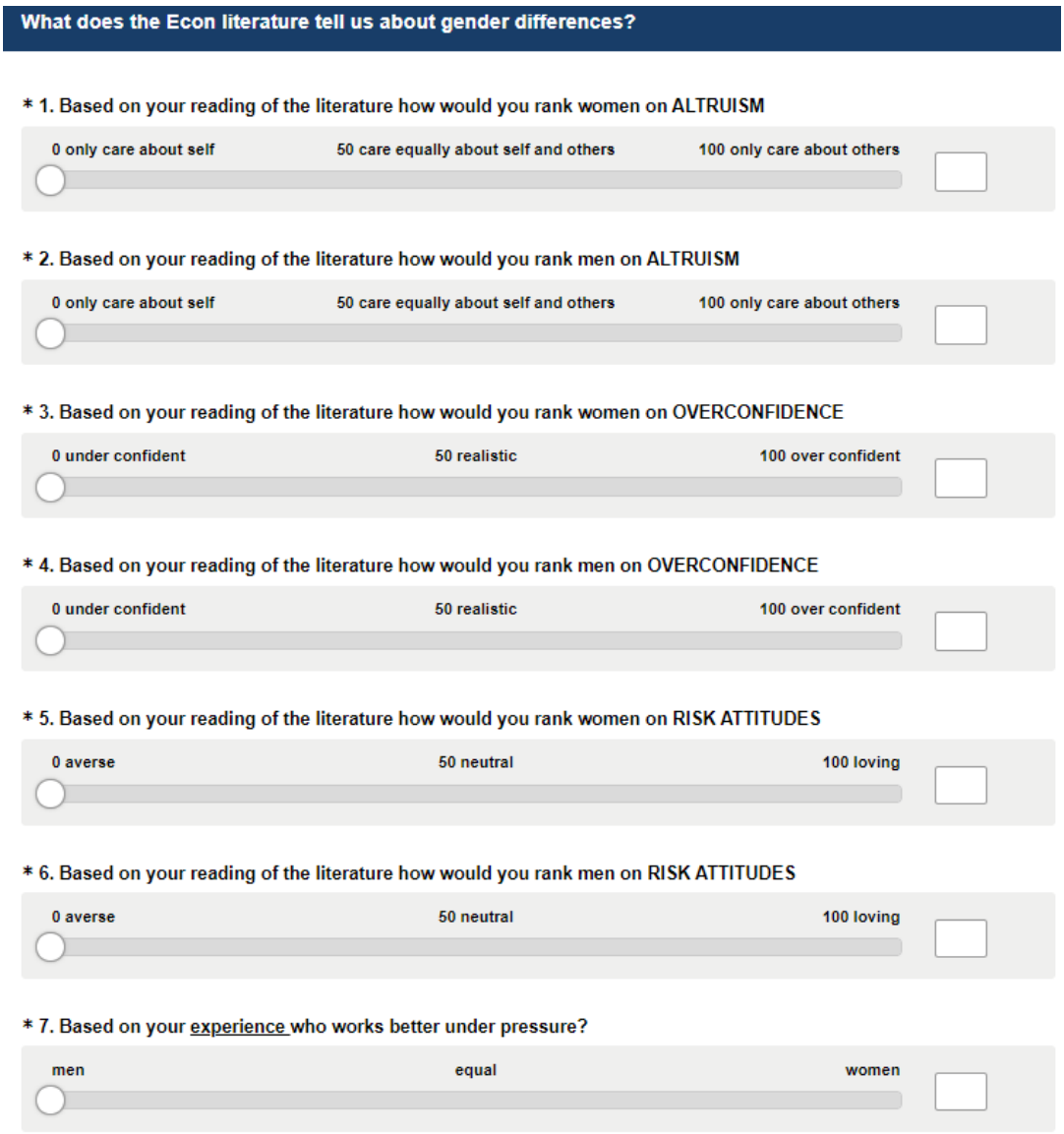


FIGURE C2: Expert survey questions 2

What does the Econ literature tell us about gender differences?

8. Tell us about yourself (tick all that applies)

	micro applied (labor, development, pf)	micro theory/io	macro	international trade	finance	Econ history	other
assistant/associate	○	○	○	○	○	○	○
full professor	○	○	○	○	○	○	○
other	○	○	○	○	○	○	○

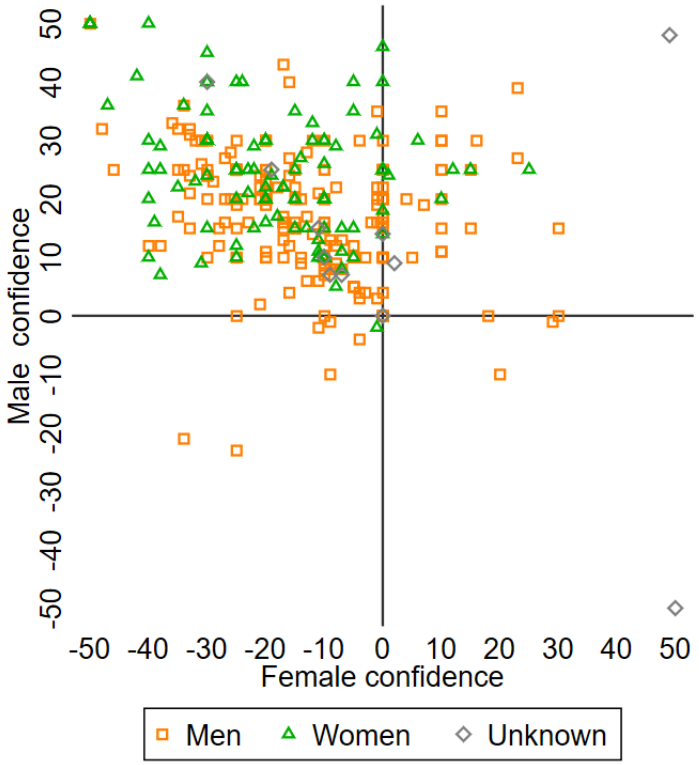
9. Tell us about yourself

	25-40	40-50	50-60	60+
male	○	○	○	○
female	○	○	○	○
other	○	○	○	○

TABLE C1: Field of specialization of survey respondents

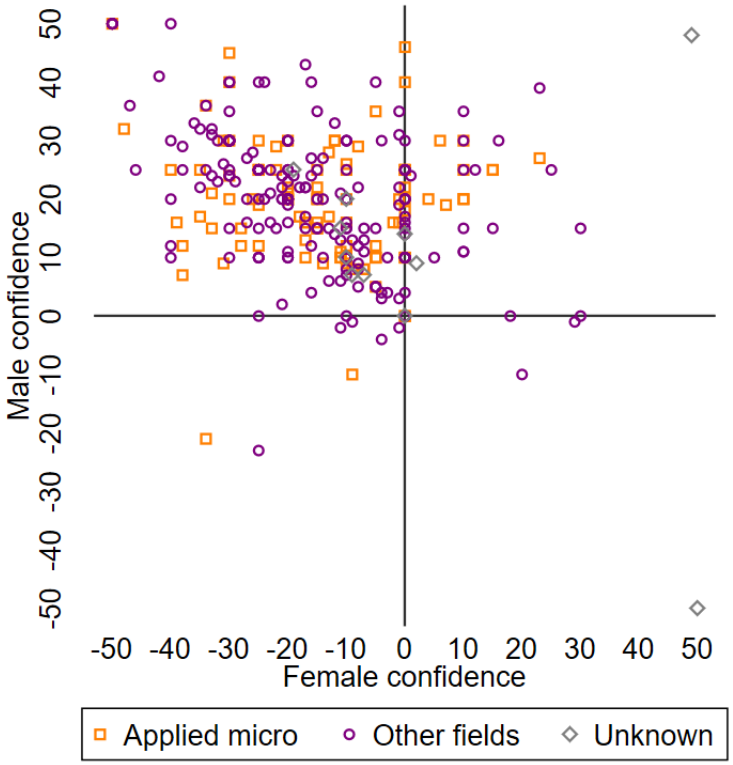
Specialization	No.
Econ History	10
Finance	39
International Trade	33
Macro	70
Micro Applied	108
Micro Theory or IO	55
Other	16

FIGURE C3: Survey results on confidence levels by gender: Men (N=220), women (N=111), and unknown (N=11)



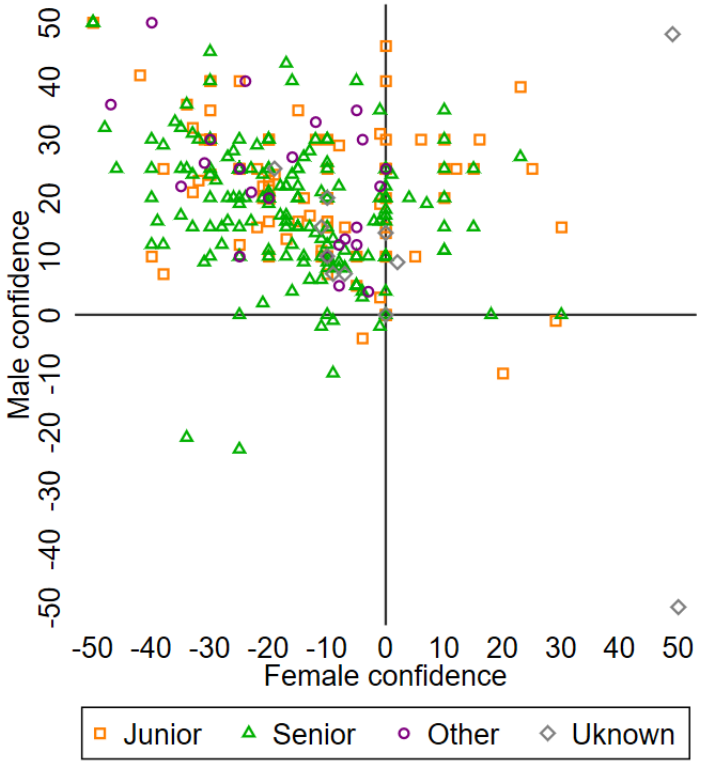
(A) Note: No respondents chose "other" as their gender. Please refer to Figures C1 and C2 for the survey questions.

FIGURE C4: Survey results on confidence levels by field: Applied micro (N=108), other fields (N=234), unknown (N=11)



(A) Note: Please refer to Figures C1 and C2 for the survey questions.

FIGURE C5: Survey results on confidence levels by seniority: Junior i.e. Assistant/Associate Professor (N=111), senior i.e. Full Professor (N=196), others (N=24), and unknown (N=11)



(A) Note: Please refer to Figures C1 and C2 for the survey questions.

C.2 Papers used in analysis

TABLE C2: List of papers used

Title	Authors and Year	Journal	Country	Exp[1]	Bhm diff[2]	Bhm levels[3]
Ask and You Shall Receive? Gender Differences in Regrades in College	Li and Zafar (2020)	IZA Discussion Paper	USA	L	Y	Y
Beliefs about Gender	Bordalo et al. (2019)	American Economic Review	USA	L	N	N
Best-of-five contest: An experiment on gender differences	Mago and Razzolini (2019)	Journal of Economic Behavior & Organization	USA	L	Y	N
Brave boys and play-it-safe girls: Gender differences in willingness to guess in a large scale natural field experiment	Iriberri and Rey-Biel (2021)	European Economic Review	Spain	F	Y	Y
Cancelling out early age gender differences in competition: an analysis of policy interventions	Sutter et al. (2016)	Experimental economics	Austria	H	N	N
Competing with confidence: The ticket to labor market success for college-educated women	Kamas and Preston (2018)	Journal of Economic Behavior & Organization	USA	L	N	N
Confidence interval estimation tasks and the economics of overconfidence	Cesarini et al. (2006)	Journal of Economic Behavior & Organization	Sweden	L	Y	N
Culture, Institutions, and the Gender Gap in Competitive Inclination: Evidence from the Communist Experiment in China	Zhang (2019)	Economic Journal	China	L	Y	Y
Do differing pay schemes help close the gender gap in overconfidence?	Lahav et al. (2015)	Economics bulletin	Israel	L	Y	Y
Do women give up competing more easily? Evidence from the lab and the Dutch Math Olympiad	Buser and Yuan (2019)	American Economic Journal: Applied Economics	Netherlands	M	N	N

Title	Authors and Year	Journal	Country	Exp[1]	Bhm diff[2]	Bhm levels[3]
Do women have more shame than men? An experiment on self-assessment and the shame of overestimating oneself	Ludwig et al. (2017)	European Economic Review	Germany and Austria	L	Y	Y
Do Women Shy Away From Competition? Do Men Compete Too Much?	Niederle and Vesterlund (2007)	Quarterly Journal of Economics	USA	L	N	N
Does affirmative action reduce gender discrimination and enhance efficiency? New experimental evidence	Beaurain and Masclet (2016)	European Economic Review	France	L	N	N
Evidence On Self-Stereotyping And The Contribution Of Ideas	Coffman (2014)	Quarterly Journal of Economics	USA	L	N	N
Gender and competition in adolescence: task matters	Dreber et al. (2014)	Experimental Economics	Sweden	L	Y	Y
Gender and confidence: are women underconfident	Jakobsson (2012)	Applied Economics Letters	Sweden	F	N	N
Gender and Self-Selection Into a Competitive Environment: Are Women More Overconfident Than Men?	Nekby et al. (2008)	Economics Letters	Sweden	F	Y	N
Gender differences in overconfidence and risk taking: Do self-selection and socialization matter?	Hardies et al. (2013)	Economics Letters	Belgium	L	N	N
Gender differences in seeking challenges: The role of institutions	Niederle and Yestrumskas (2008)	NBER Working Paper	USA	L	N	N
Gender, Competition and the Efficiency of Policy Interventions	Balafoutas and Sutter (2010)	IZA Discussion Paper	Austria	L	N	N
Gender, competitiveness, and career choices	Buser et al. (2014)	Quarterly Journal of Economics	Netherlands	L	N	N
How Costly is Diversity? Affirmative Action in Light of Gender Differences in Competitiveness	Niederle et al. (2013)	Management Science	USA	L	N	N
Image and Misreporting	Ewers and Zimmermann (2015)	Journal of the European Economic Association	Germany	L	N	N
Is financial instability male driven? Gender and cognitive skills in experimental asset markets	Cueva and Rustichini (2015)	Journal of Economic Behavior & Organization	UK	L	Y	N

Title	Authors and Year	Journal	Country	Exp[1]	Bhm diff[2]	Bhm levels[3]
Is there a gender Gap in Preschoolers' Competitiveness? An experiment in the U.S	Samak (2013)	Journal of Economic Behavior & Organization	USA	L	N	N
Judgemental Overconfidence, Self-Monitoring, and Trading Performance in an experimental Financial Market	Biais et al. (2005)	Review of Economic Studies	France and UK	L	Y	N
Laboratory Evidence on the Effects of Sponsorship on the Competitive Preferences of Men and Women	Baldiga and Coffman (2018)	Management Science	USA	L	N	N
Male And Female Competitive Behavior–Experimental Evidence	Gupta et al. (2005)	IZA Discussion Paper	France	L	N	N
No Gender Difference in Willingness to Compete When Competing against Self	Apicella et al. (2017)	American Economic Review	USA, Online	M	Y	Y
Overconfidence as a social bias: Experimental evidence	Proeger and Meub (2014)	Economics Letters	Germany	L	N	N
Overconfidence and investment: An experimental approach	Pikulina et al. (2017)	Journal of Corporate Finance	Netherlands	L	N	N
Preferences And Biases In Educational Choices And Labour Market Expectations: Shrinking The Black Box Of Gender	Reuben et al. (2017)	Economic Journal	USA	L	N	N
Self-confidence and strategic behavior	Charness et al. (2018)	Experimental Economics	Netherlands	L	Y	N
The emergence of male leadership in competitive environments	Reuben et al. (2012)	Journal of Economic Behavior & Organization	USA	L	N	N
The gender gap in self-promotion	Exley and Kessler (2019)	NBER Working Paper	USA	L	Y	N
The importance of being confident; gender, career choice, and willingness to compete	Kamas and Preston (2012)	Journal of Economic Behavior & Organization	USA	L	Y	Y
The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices	Wozniak et al. (2014)	Journal of Labor Economics	USA	L	N	N
Willingness to Compete: Family Matters	Almås et al. (2016)	Management Science	Norway	L	Y	Y

Notes:
1. L=Lab experiment, F=Field Experiment, H=Hybrid experiment, M=Mix of experiments in a single paper.
2. Whether the results from the paper are used in the BHM for gender differences in confidence (Y=Yes, N=No).
3. Whether the results from the paper are used in the BHM for gender-specific overconfidence (Y=Yes, N=No).

C.3 Bayesian results for alternative sub-samples

FIGURE C6: Model comparison - overconfidence by gender, extensive margin sample

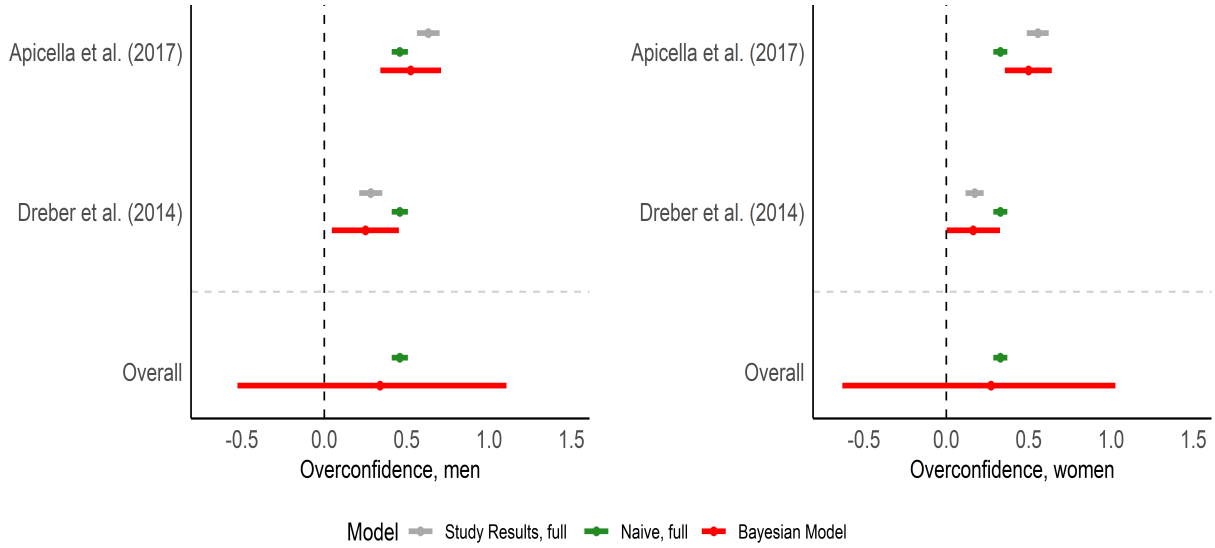


FIGURE C7: Model comparison - overconfidence by gender, intensive margin sample

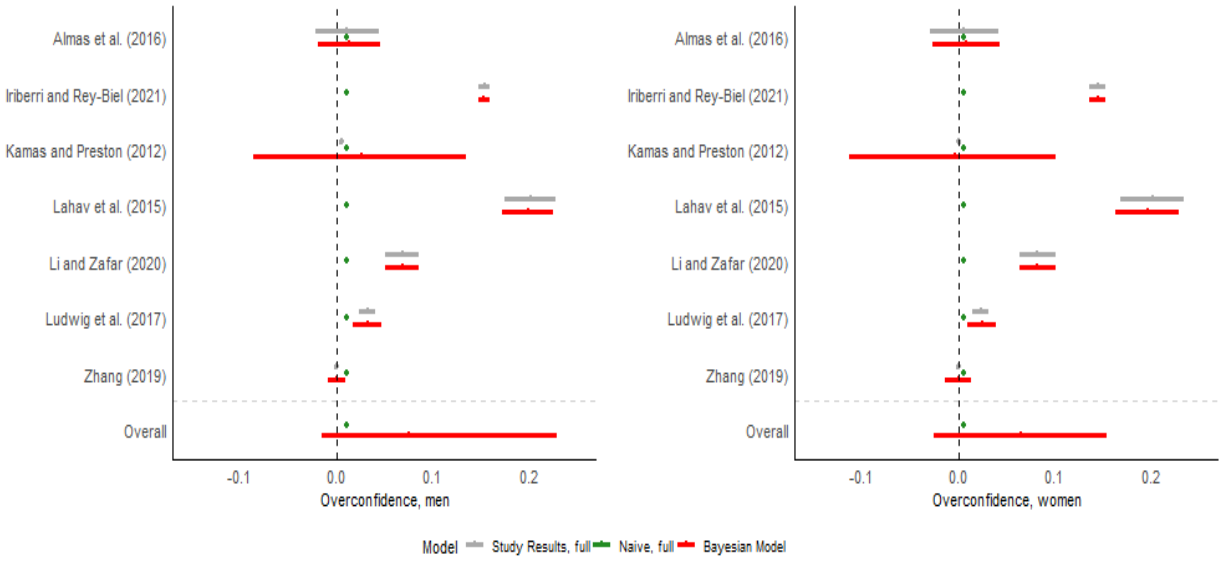


FIGURE C8: Model comparison - gender differences in overconfidence, extensive margin sample

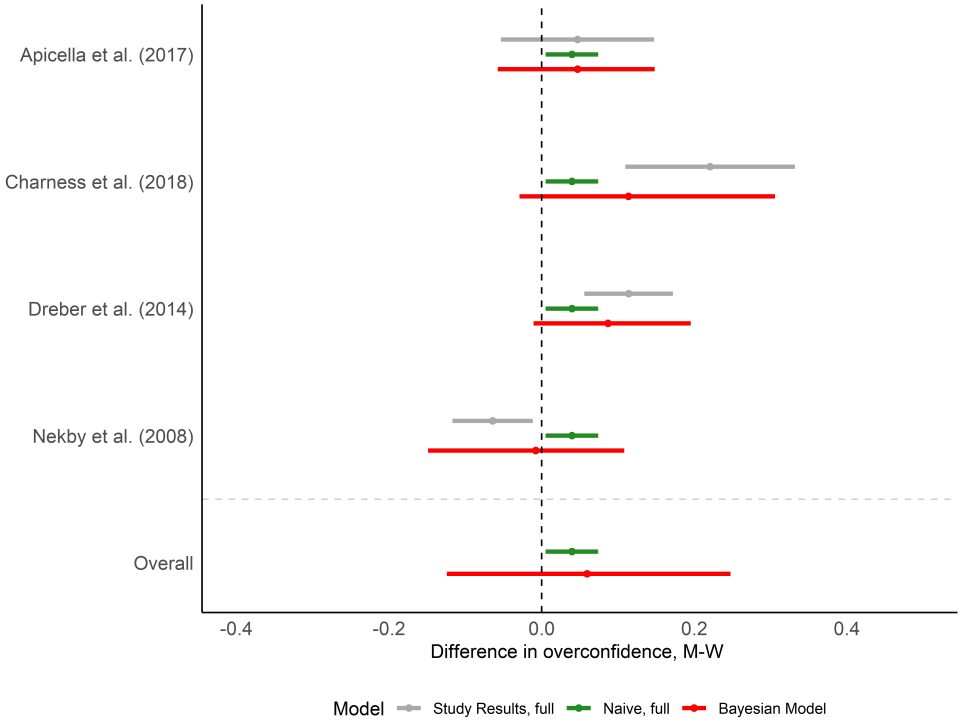
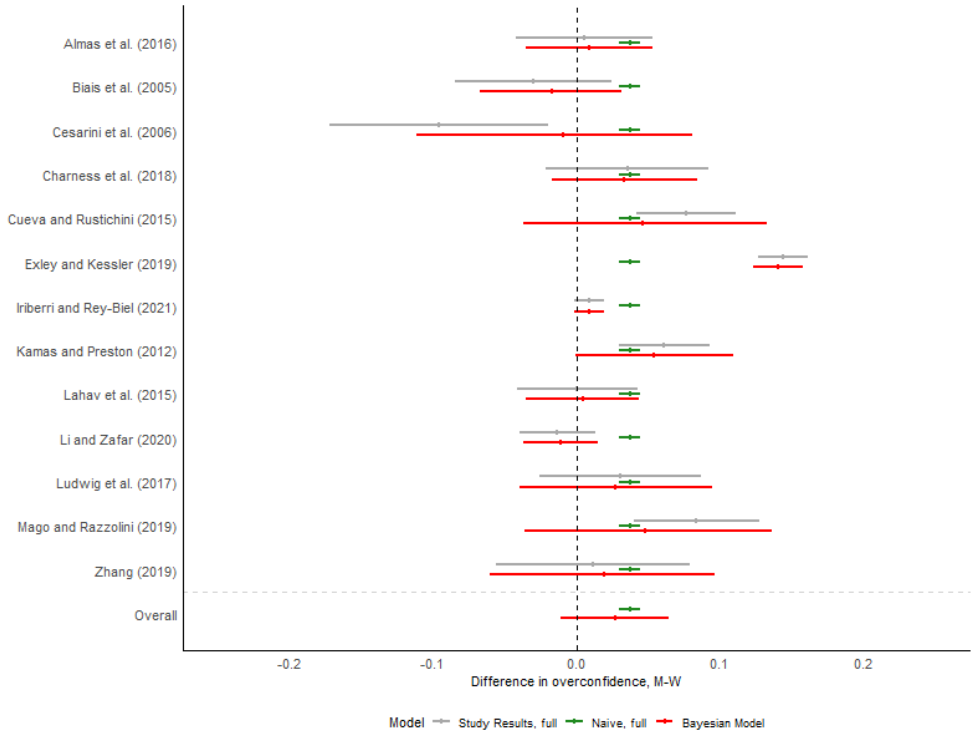


FIGURE C9: Model comparison - gender differences in overconfidence, intensive margin sample



C.4 Robustness analysis

TABLE C3: Alternative functional forms on priors: overconfidence, men

Model Priors	$\hat{\beta}^p$	$\hat{\sigma}^p$	2.5%	50%	97.5%
$\beta \sim$normal(0,1); $\sigma \sim$normal(0,1)	0.3915	0.1503	0.0855	0.3935	0.6868
$\beta \sim$cauchy(0,1); $\sigma \sim$normal(0,1)	0.3853	0.1492	0.0810	0.3869	0.6792
$\beta \sim$normal(0,10); $\sigma \sim$normal(0,1)	0.3999	0.1538	0.0890	0.4007	0.7065
$\beta \sim$cauchy(0,10); $\sigma \sim$normal(0,1)	0.3999	0.1537	0.0885	0.4006	0.7074
$\beta \sim$normal(0,1); $\sigma \sim$normal(0,10)	0.3906	0.1551	0.0769	0.3923	0.6954
$\beta \sim$cauchy(0,1); $\sigma \sim$normal(0,10)	0.3837	0.1540	0.0680	0.3864	0.6850
$\beta \sim$normal(0,10); $\sigma \sim$normal(0,10)	0.3952	0.1681	0.0613	0.3993	0.7107
$\beta \sim$cauchy(0,10); $\sigma \sim$normal(0,10)	0.4015	0.1568	0.0884	0.4017	0.7176
$\beta \sim$normal(0,1); $\sigma \sim$uniform(0,1)	0.3907	0.1524	0.0793	0.3924	0.6932
$\beta \sim$cauchy(0,1); $\sigma \sim$uniform(0,1)	0.3854	0.1518	0.0746	0.3874	0.6843
$\beta \sim$normal(0,10); $\sigma \sim$uniform(0,10)	0.3993	0.1575	0.0832	0.3996	0.7147
$\beta \sim$cauchy(0,10); $\sigma \sim$uniform(0,10)	0.4000	0.1563	0.0854	0.4004	0.7123

TABLE C4: Alternative functional forms on priors: overconfidence, women

Model Priors	$\hat{\beta}^p$	$\hat{\sigma}^p$	2.5%	50%	97.5%
$\beta \sim$normal(0,1); $\sigma \sim$normal(0,1)	0.3524	0.1541	0.0382	0.3538	0.6570
$\beta \sim$cauchy(0,1); $\sigma \sim$normal(0,1)	0.3472	0.1526	0.0357	0.3490	0.6474
$\beta \sim$normal(0,10); $\sigma \sim$normal(0,1)	0.3601	0.1562	0.0444	0.3604	0.6712
$\beta \sim$cauchy(0,10); $\sigma \sim$normal(0,1)	0.3617	0.1557	0.0499	0.3621	0.6727
$\beta \sim$normal(0,1); $\sigma \sim$normal(0,10)	0.3906	0.1551	0.0769	0.3923	0.6954
$\beta \sim$cauchy(0,1); $\sigma \sim$normal(0,10)	0.3837	0.1540	0.0680	0.3864	0.6850
$\beta \sim$normal(0,10); $\sigma \sim$normal(0,10)	0.3952	0.1681	0.0613	0.3993	0.7107
$\beta \sim$cauchy(0,10); $\sigma \sim$normal(0,10)	0.4015	0.1568	0.0884	0.4017	0.7176
$\beta \sim$normal(0,1); $\sigma \sim$uniform(0,1)	0.3907	0.1524	0.0793	0.3924	0.6932
$\beta \sim$cauchy(0,1); $\sigma \sim$uniform(0,1)	0.3854	0.1518	0.0746	0.3874	0.6843
$\beta \sim$normal(0,10); $\sigma \sim$uniform(0,10)	0.3993	0.1575	0.0832	0.3996	0.7147
$\beta \sim$cauchy(0,10); $\sigma \sim$uniform(0,10)	0.3341	0.1394	0.0472	0.3371	0.6044

TABLE C5: Alternative functional forms on priors: gender differences in overconfidence

Model Priors	$\hat{\beta}^p$	$\hat{\sigma}^p$	2.5%	50%	97.5%
$\beta \sim$normal(0,1); $\sigma \sim$normal(0,1)	0.0936	0.0542	-0.0155	0.0942	0.1995
$\beta \sim$cauchy(0,1); $\sigma \sim$normal(0,1)	0.0932	0.0542	-0.0156	0.0939	0.1988
$\beta \sim$normal(0,10); $\sigma \sim$normal(0,1)	0.0938	0.0543	-0.0153	0.0944	0.1997
$\beta \sim$cauchy(0,10); $\sigma \sim$normal(0,1)	0.0937	0.0543	-0.0157	0.0944	0.1997
$\beta \sim$normal(0,1); $\sigma \sim$normal(0,10)	0.0941	0.0543	-0.0149	0.0946	0.2003
$\beta \sim$cauchy(0,1); $\sigma \sim$normal(0,10)	0.0928	0.0541	-0.0159	0.0933	0.1981
$\beta \sim$normal(0,10); $\sigma \sim$normal(0,10)	0.0939	0.0545	-0.0157	0.0945	0.2001
$\beta \sim$cauchy(0,10); $\sigma \sim$normal(0,10)	0.0936	0.0543	-0.0154	0.0941	0.1996
$\beta \sim$normal(0,1); $\sigma \sim$uniform(0,1)	0.0937	0.0544	-0.0159	0.0942	0.2001
$\beta \sim$cauchy(0,1); $\sigma \sim$uniform(0,1)	0.0931	0.0539	-0.0152	0.0937	0.1988
$\beta \sim$normal(0,10); $\sigma \sim$uniform(0,10)	0.0938	0.0542	-0.0151	0.0944	0.1996
$\beta \sim$cauchy(0,10); $\sigma \sim$uniform(0,10)	0.0938	0.0544	-0.0154	0.0944	0.2001

TABLE C6: Rubin model: posterior-mean of overconfidence of men and women across subsamples

Sample	N	J	$\hat{\beta}^p$	$\hat{\sigma}^p$	percentiles				
					2.5	25	50	75	97.5
Extensive margin sample, men	7	2	0.512	0.098	0.313	0.455	0.512	0.569	0.707
Extensive margin sample, women	7	2	0.442	0.108	0.223	0.379	0.442	0.506	0.662
Intensive margin sample, men	17	7	0.037	0.019	-0.000	0.025	0.037	0.050	0.075
Intensive margin sample, women	17	7	0.027	0.020	-0.012	0.014	0.027	0.040	0.066
Full sample, men	24	9	0.409	0.091	0.229	0.349	0.409	0.469	0.591
Full sample, women	24	9	0.315	0.085	0.147	0.259	0.315	0.371	0.484

TABLE C7: Rubin model: posterior-mean of gender differences in overconfidence across subsamples

Sample	N	J	$\hat{\beta}^p$	$\hat{\sigma}^p$	percentiles				
					2.5	25	50	75	97.5
Extensive margin sample	13	4	0.074	0.039	-0.002	0.048	0.073	0.098	0.154
Intensive margin sample	26	13	0.034	0.013	0.007	0.026	0.034	0.043	0.060
Full sample	39	16	0.127	0.037	0.054	0.103	0.127	0.151	0.202

TABLE C8: Rubin model: posterior-mean of gender differences in overconfidence with citation-weighted s.e.

Sample	Data	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	percentiles				
					2.5	25	50	75	97.5
Extensive margin sample	Original s.e.	13	0.074	0.039	-0.002	0.048	0.073	0.098	0.154
Extensive margin sample	Citation adjusted s.e.	13	0.080	0.041	-0.001	0.054	0.079	0.105	0.162
Intensive margin sample	Original s.e.	26	0.034	0.013	0.007	0.026	0.034	0.043	0.060
Intensive margin sample	Citation adjusted s.e.	26	0.029	0.018	-0.006	0.017	0.029	0.040	0.064

TABLE C9: Rubin model: posterior mean of gender differences in overconfidence with alternative priors

Prior on β	N	$\hat{\beta}^p$	$\hat{\sigma}^p$	percentiles				
				2.5	25	50	75	97.5
$\beta \sim N(0, 1)$	39	0.127	0.037	0.054	0.103	0.127	0.151	0.202
$\beta \sim N(0, 0.078^2)$	39	0.003	0.006	0.061	0.106	0.130	0.153	0.203
$\beta \sim N(1.46, 0.078^2)$	39	1.463	0.006	1.451	1.459	1.463	1.467	1.475
$\beta \sim N(1.46, 1)$	39	0.129	0.035	0.058	0.106	0.130	0.152	0.201