Program Evaluations and Policy Spending *

See latest version **here**

Michelle Rao

Last Updated: November 12, 2024

Abstract

Program evaluations are motivated in part by a desire to improve the effectiveness of policy spending. 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 evaluated programs, I find a precise 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 which are conducted quickly (within four years of the effect year) are significantly predictive of spending. Thus, timeliness may be an overlooked aspect of the evidence-to-policy pipeline.

*I am grateful to my advisers, Oriana Bandiera, Gharad Bryan, Robin Burgess, and Rachael Meager for their continued support on this project. For helpful comments and discussions, I thank Tim Besley, Stefan Dercon, Gabriela Deschamps, Nilmini Herath, Amen Jalal, Gilat Levy, Santiago Levy, Gabriel Leite Mariante, Canishk Naik, Ashley Pople, Maarten De Ridder, Jack Thiemel, Sarah Winton, Liam Wren-Lewis, and seminar and conference participants at LSE, OxDEV, and Doctorissimes. Excellent research assistance was provided by Miguel Gonzalez Lugo. Funding provided by STICERD is gratefully acknowledged. Contact: Department of Economics, London School of Economics and Political Science, m.rao3@lse.ac.uk

1 Introduction

Program evaluations are increasingly integrated into policy, 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, the empirical relationship between the results of these evaluations and key decisions, such as policy spending, has not yet been rigorously studied. Program evaluations give credible estimates of impact and can, in theory, improve the efficiency of policy spending [Duflo and Kremer, 2003]. Yet, their applicability to policy decisions can also be constrained 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 - and the features of evidence that matter - is a fundamental step to maximising the policy impact of research.

This paper contributes to this understanding by studying program evaluations and policy spending of Conditional Cash Transfers (CCTs) in Latin America and the Caribbean. Due to their proliferation and heavy evaluation, CCTs are often cited as a success story for evidence-based policy spending [e.g Duflo and Banerjee, 2011, Angrist and Pischke, 2010]. The early studies of Mexico's PROGRESA [e.g. Gertler, 2004, Schultz, 2004] contributed to a culture of evaluation of CCTs particularly for countries in Latin America and the Caribbean [Fiszbein and Schady, 2009]. Indeed, between 2000 to 2015, there were 31 evaluated CCTs across 17 countries in the region, often in collaboration with government. However, little is known about the relationship between CCT evaluation outcomes and policy spending.

This is the focus of my study. I examine the relationship between program evaluation outcomes and spending on evaluated 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, representing 468 headline results on the causal impact of CCTs on poverty-related outcomes. Using this data, I examine patterns of policy spending that are consistent with both, immediate and gradual patterns of evidence-use.

To study this relationship empirically, I need to clearly define what patterns in the data would be consistent with evidence-use. Using a simple theoretical model of policymaking under uncertainty, I show that under basic assumptions on evidence-quality, evidence-use would be reflected empirically by a positive relationship between evalua-

tion outcomes and spending if: (1) policymakers use evidence to update their beliefs; and (2) the perceived benefits of adjusting policy spending outweigh the political costs. The relationship between research findings and spending therefore depends not only on the evidence-base, but also on its interaction with political or other constraints.

There are two challenges to discerning this relationship in the data. First, even if policymakers are using evidence, I cannot observe the subset of evaluations – the information set – that policymakers use to make spending decisions. CCT programs are repeatedly evaluated. 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 information policymakers extract, and how they summarise the information. Thus, studying the systematic 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 impact across two types information sets: individual evaluations, and cumulative evidence from each country. First, using variation in findings across individual evaluations, I find a precise zero association between evaluation outcomes and policy spending. This relationship 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. Second, using variation in the cumulative evidence-base across countries, I again find a zero relationship between a country's aggregate evidence and policy spending. As the zero relationship holds for the subset of evaluations that are conducted in collaboration with policymakers, these findings are unlikely to be driven by lack of policy awareness. Rather, they suggest that outcomes from evaluations do not overcome the constraints to evidence-use. In the final part of the paper, I explore this possibility using variation in evaluation characteristics associated with lower constraints. I find that larger evaluation outcomes from timely evaluations, i.e. evaluations that are available faster relative to the effect year, are associated with larger changes in policy spending.

These findings have implications for optimal research for policy impact. While studies of optimal research design often assume 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 cannot be taken as given. Rather, the positive

association between evaluation outcomes and changes in spending for timely evaluations is suggestive of the presence of constraints, that may be increasing over time.[1] Thus, increasing the timeliness of evaluation may be an overlooked aspect of the evidence-to-policy pipeline.

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 holds regardless of the way in which I aggregate reported treatment effects from each evaluation. There is no association between spending and the statistical significance of program evaluation results, as captured by the mean or maximum of the precision-weighted treatment effect (i.e. the t-statistic). There is also no association between spending and the magnitude of program evaluation results, as captured by the mean or maximum effect size of headline results. The estimated relationship between treatment effects and 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 evaluations 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 relative to these evidence-based priors do not correspond with larger changes in spending. Evaluations that are more positive, relative to the existing evidence base do not correspond with larger increases in spending. Furthermore, evaluations that are more negative, relative to the existing evidence base, do not correspond with larger decreases in spending.[2] These results hold regardless of assumptions on the weight countries place on evidence from other

[1] This could be explained, for instance, by increasing political costs, and decreasing external validity of treatment effects [e.g. Rosenzweig and Udry, 2020] over time.

[2] 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.

countries when forming prior beliefs – that is, assumptions on the perceived external validity of evaluations from other countries.

Beyond treatment effects, the strength of evaluation results are 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 are. Authors use more positive language to describe their research findings when there are larger treatment effects. However, I find that more positively framed evaluations do not correspond with 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 precise zero relationship between individual evaluations and program spending, sophisticated users of evidence may instead learn from the aggregate evidence-base. This is justified in a setting with low external validity and hence limited scope for learning from individual evaluations [e.g. Allcott, 2015, Rosenzweig and Udry, 2020].[3] 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. 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 causal estimates of impact and subsequent spending. These results suggest that either policymakers do not adjust their spending in response to impact es-

[3] 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. Rosenzweig and Udry [2020] find that in the presence of aggregate shocks, internally valid findings do not generalize across time periods, even to the same population of interest.

timates; 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 observe changes in beliefs. However, given that policymakers are highly trained, and are often directly or indirectly involved in the evaluation of CCTs, these results seem unlikely to be driven by a lack of policy awareness. Rather, they suggest that program evaluations do not overcome the practical or political constraints to evidence-use.

I examine the role of constraints by considering patterns of responsiveness to subsets of evidence that are likely to be more policy relevant. I consider 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 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 results are available faster than the mean of four years after the effect year, there is a positive relationship between paper-level findings and spending. This positive association is driven by timely evaluations available in years with low political constraints. In particular, the association between policy spending and research results is highest when the political party in power is the same at the effect year, and the first date of publication.

Most closely related to my study are DellaVigna et al. [2022] and Wang and Yang [2021], who study policy experimentation and evidence-use in government institutions.

Similar to DellaVigna et al. [2022], 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 adoption.

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., 2022], 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], and 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 2 describes the data and context. Section 3 lays out the conceptual framework and empirical strategy. Section 4 and section 5 outlines the main results on paper-level findings and spending, and cumula-

tive findings and spending, respectively. Section 6 discusses the interpretation of the average zero relationship, and implications for the evidence to policy pipeline. Section 7 explores heterogeneity in responses to different features of evidence. Section 8 concludes.

2 Data & context

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 2.1 and 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 2.3, I provide some descriptive facts and context about evaluations and spending on CCTs.

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. 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[4]. 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 analysis.

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 [5]. I obtain

[4]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.

[5]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

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 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. The remaining 79 use non-experimental methods, such as propensity score matching, Differences in Difference, Regression Discontinuity and Instrumental Variables.

Table 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.

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

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.

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 measure 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[6]. Using this information, I identify whether or not any of the study authors collaborated with the government at some point during program evaluation.[7] I find that 65 out of 128 studies in my sample have at least one author affiliated with the governing institution. 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.

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.

[6]See Appendix B.2 for more detail on data collection of study characteristics

[7]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.

2.2 Program characteristics

I map the program evaluations of Conditional Cash Transfers to data on annual programme 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.

2.3 Descriptive facts about CCT program evaluations & spending

Conditional Cash Transfer programs are a widespread policy instrument for social protection and are heavily studied, particularly across Latin America and the Caribbean. One of most renowned CCT programs is Mexico's PROGRESA, a program that provided cash transfers to poor rural households, conditional on education, health and nutritional activities. A defining feature of PROGRESA was its rigorous evaluation. From the inception of the program, policymakers established a research partnership with the International Food Policy Research Institute to evaluate the causal impact of the program on the targeted population.

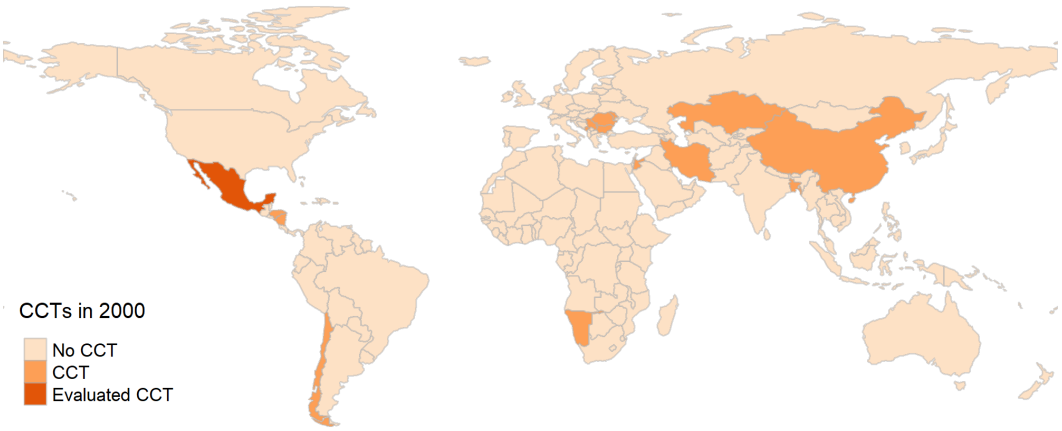
The evaluation of PROGRESA created a trend of rigorous evaluation of Conditional Cash Transfer programs, particularly in Latin America and the Caribbean [Fiszbein and Schady, 2009]. This is reflected in figure 1. By 2015, Conditional Cash Transfers had spread throughout low and middle income countries as an effective policy innovation for poverty reduction. However, systematic evaluation is particularly evident in Latin America and the Caribbean, where almost all countries in the region has an established CCT program with an associated program evaluation by 2015.

These patterns are even more striking when considering the patterns in my data. In 2015, almost all countries in Latin America and the Caribbean had an active CCT program (Figure 2a). 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.

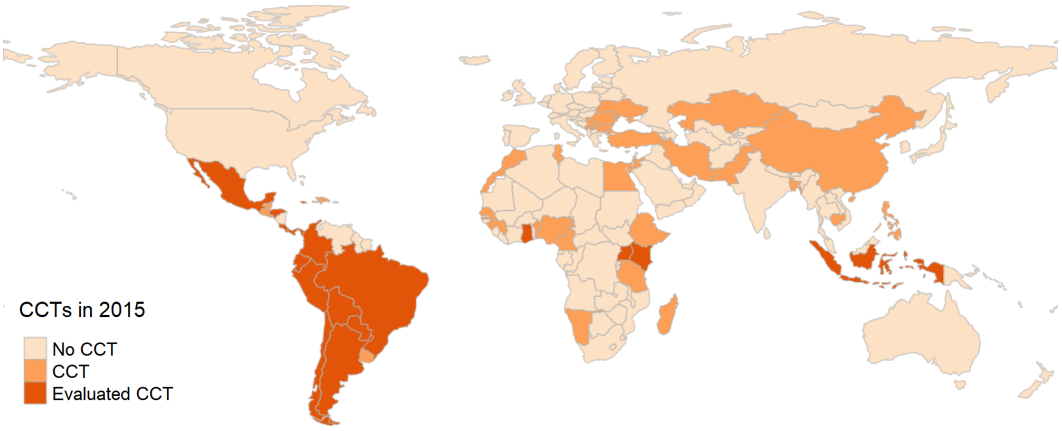
Alongside the expansion in spending, CCT programs in the region are repeatedly evaluated over time. I identify 128 program evaluations estimating the causal impact of 31 CCT programs across 17 countries. As seen in Figure 2b, while Mexico's PROGRESA/ Oportunidades is by far the most heavily studied program, evaluations are common and widespread. The median country has had seven causal evaluations on the impact of CCTs on poverty-related outcomes.

These evaluations are highly embedded in government, suggesting that policymakers are likely to be aware of evidence base (table 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 1. The size of the treatment effects varies across studies, but most countries in my sample have experienced



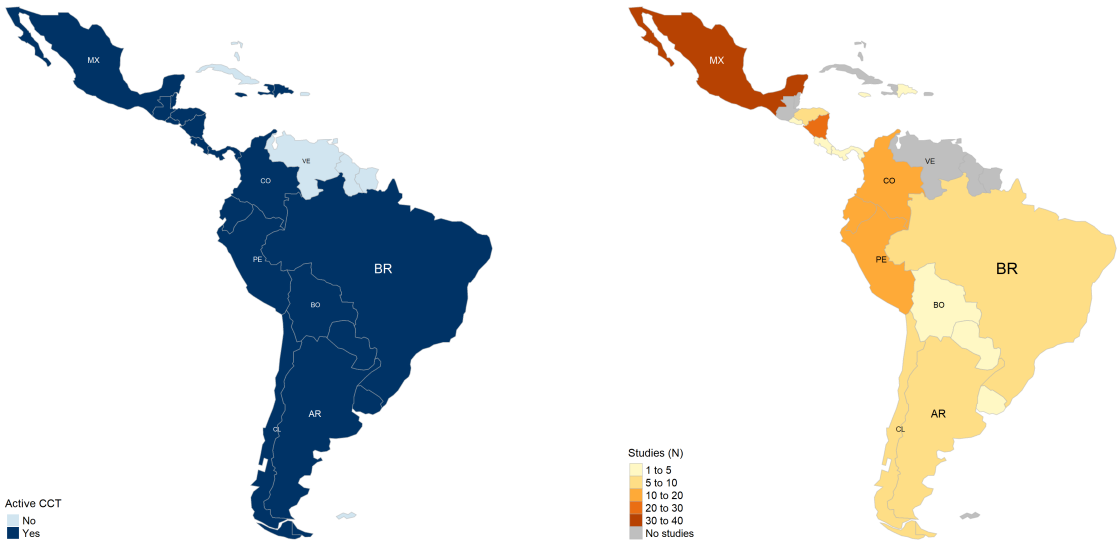
(a) 2000



(b) 2015

Figure 1: 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) Active Conditional Cash Transfers

(b) N. program evaluations by country

Figure 2: Active cash transfers and cumulative program evaluations in 2015

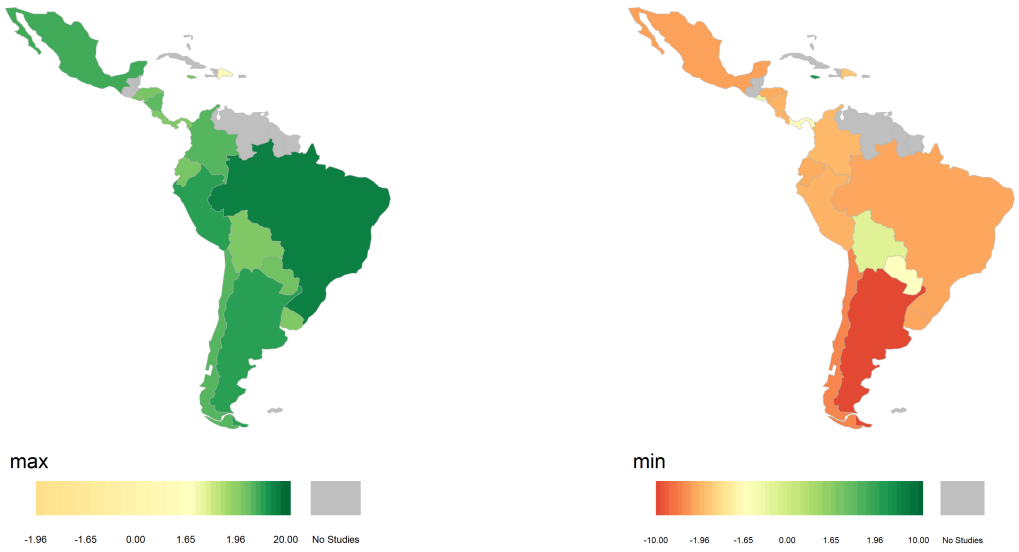
Notes: Active CCTs and number of aggregate program evaluations on CCTs by country in Latin America and the Caribbean in 2015.

Table 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-gov link: 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.

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



(a) Maximum treatment effect across all studies

(b) Minimum treatment effect across all studies

Figure 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.

3 Conceptual framework & method

To set ideas, I present a simple conceptual framework of policy spending, based on Buera et al. [2011], who study a policymaker's decision to implement market-oriented policies based on their own and neighbours' past experiences. I use the same set-up as Buera et al. [2011], but adapt the learning environment to incorporate policymaking using information signals from program evaluations.

Assume that the policymaker is benevolent and derives utility from minimising the sum of poverty, Y_{it}, and political costs, K_{it}, subject to their beliefs of how poverty

changes over time.[8] Policymakers choose θ_{it}, an indicator variable for whether or not to increase spending on a CCT program, to maximise their expected utility.[9]

The optimisation problem is thus summarised as follows:

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

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 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. 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 cost, K_{it}. Given their beliefs of γ_i, they then decide whether or not to increase spending.

The optimal policy decision is therefore given by:

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

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 implementation 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

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

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

evidence-based beliefs – that is, beliefs consistent with evidence from program evaluations. I define evidence-based beliefs as the following:

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

where $\hat{\mu}_{it-1}$ is a vector of causal estimates of impact from program evaluations on country i, with evaluation results available up to year $t-1$; and $f(.)$ is an increasing function of $\hat{\mu}_{it-1}$.

Combining equations 2 and 3, this implies that an evidence-based policymaker will increase policy spending 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 perceived impact of the policy is higher than K_{it}, the constraints to increasing spending. The relationship between causal estimates of impact and policy spending therefore depends not only on the evidence-base, but also on the interaction between features of evidence and political constraints.

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 program spending 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.

On the other hand, policymakers face several barriers to evidence-use. First, even if policymakers are inclined to use evidence, program evaluations are imperfect because they do not necessarily signal 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, even if policymakers learn from evidence, the expected benefits of changing spending will not necessarily overcome the political constraints. 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 political 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.[10]

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

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:

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

where $\hat{\mu}_{it-1}$ is the perceived causal impact 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 4 is in estimating $f(\hat{\mu}_{it})$, the perceived causal impact of a CCT program based on a given evaluation. While $f(\hat{\mu}_{it})$ is known to the policymaker, it is unobserved by the econometrician. This is due to two main reasons:

- The econometrician cannot observe 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. In other words, the **aggregation method** is also unobserved.

I thus estimate equation 4 by constructing estimates of $f(\hat{\mu}_{it})$ – summary metrics of impact – across different information sets and aggregation methods, which together, mirror different models of evidence use. 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 cumulative 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 3.1 and 3.2.

3.1 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:

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

where $\hat{\mu}_{ist-1}$ is the aggregated paper-level finding from a paper 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 categories of aggregations of findings from each individual evaluation. Each of these aggregation methods provides a different estimate of $\hat{\mu}_{ist-1}$.

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 significance of headline results. I capture the magnitude of the causal impact of CCTs by effect size, calculated as the estimated treatment effect divided by the standard deviation [11]; and the significance of research results by the test-statistic, calculated as the estimated treatment effect divided by the standard error.

An illustrating example: To understand the set of relevant signals attached to program evaluations, 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,

[11] 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.

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 paper (see table 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.

Table 3: Example of study level summary metrics based on Galiani and McEwan [2013]

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

Evaluation results, relative to the existing knowledge base: As documented in section 2, CCTs are often evaluated repeatedly over time. The median country in my sample is evaluated 7 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 existing prior beliefs.

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

$$\hat{\mu}_{ist} = \tau_{ist} - \hat{\nu}_{it} \tag{6}$$

where $\hat{\nu}_{it}$ 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. $\tau_{ist} - \hat{\nu}_{it}$ is therefore a measure of how 'surprising' a paper is, relative to the existing evidence base. $\tau_{ist} - \hat{\nu}_{it} > 0$ means that the CCT is performing better than would be expected; whereas $\tau_{ist} - \hat{\nu}_{it} < 0$ means that the CCT is underperforming, relative to expectations.

To estimate $\hat{\nu}_{it}$, 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 $\hat{\nu}_{it}$ 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:

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

$\forall s \in t$ 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 $\tau_{ist} - \hat{\nu}_{it}$ across values of $\lambda \in [0, 1]$, based on 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 instead be responsive to how research results are described and communicated. In describing study findings, researchers convey their attitudes towards the policies through language. This, in turn, can affect the beliefs and decision-making of the 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 sentiment from paper abstracts to make policy conclusions.[12]

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

[12]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.

sentiment for each paper is defined as:

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

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

3.2 Aggregating a country's evidence-base

What if policymakers are responding to the cumulative body of evidence? 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. Changing spending in line with the evidence may also take time, due to institutional and political constraints to policy change. As a result, evidence-based policy spending may be reflected through patterns in aggregate spending and cumulative bodies of evidence over time.

I therefore consider the relationship between cumulative bodies of evidence and spending, as follows:

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

where $\hat{\mu}_i$ is the estimated posterior mean of findings from all CCTs evaluations conducted on country i; 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. 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. In the second stage of the estimation, I use the posterior estimates of evaluation-level findings from the first stage to estimate a country-level posterior mean of the cumulative evidence base.

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. For each evaluation j, I estimate the posterior mean of the evaluation, $\hat{\tau}_{ji}$, as:

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

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:

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

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).

4 Individual evaluations & spending

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%. [13]

In figure 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.

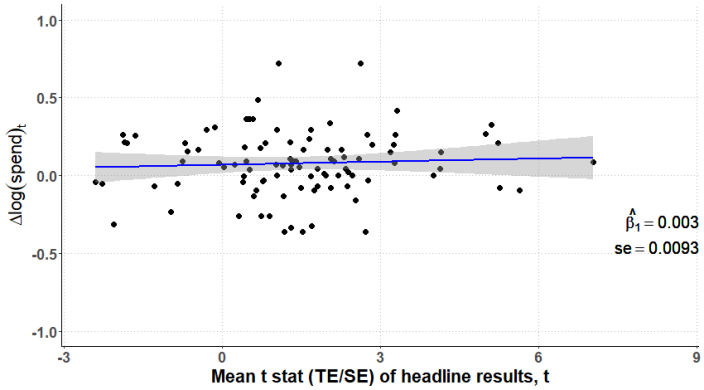


Figure 4: Mean t statistic, and changes in spending

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.

This zero correlation is not driven by choices in the aggregation of reported treatment effects. In figure 5, I plot the relationship between causal estimates of program impact and spending across four summary metrics of headline results: the maximum magnitude, the mean magnitude, the maximum statistical significance, and the mean

[13] 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.

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.

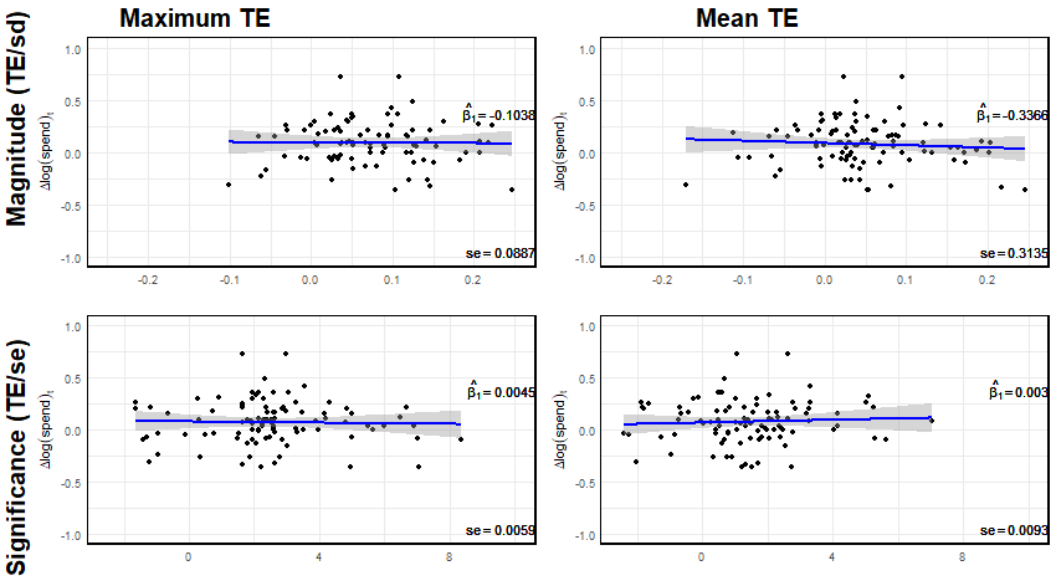


Figure 5: Treatment effects and changes in spending on the same program, across measures of evaluation outcomes

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).

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

This 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?

Table 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 .

The policymaker's consumption of evidence is unobserved. However, I can proxy for policymaker awareness using information on institutional demand, and government-author relationships.

In figure 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.

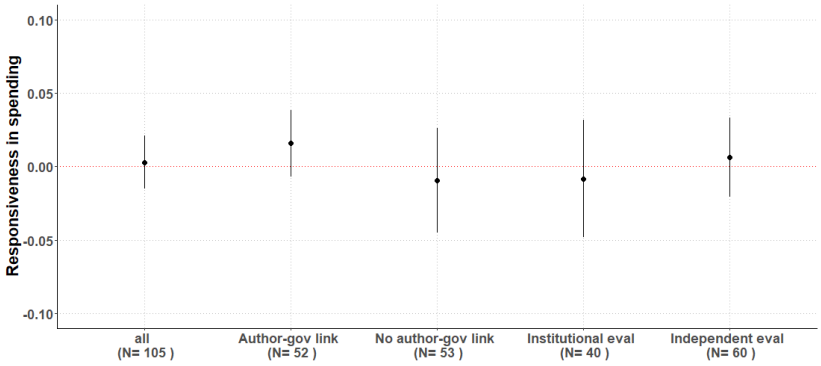


Figure 6: Responsiveness in spending to subsets of evidence, by government-author relationships and source of evaluation

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.

How do these results relate to organisational or political constraints? Even if policymakers are aware of evaluation outcomes, and use evidence to update their beliefs, this would only translate to changes in spending if policymakers are able to overcome the constraints to policy change. Evaluation results made available in years with lower political or organisational constraints to policy change may therefore be associated with higher responsiveness to treatment effects.

To examine the role of organisational constraints, 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.

I explore the role of political constraints in figure 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. Moreover, countries with functioning democracies would be more able to hold politicians' accountable – therefore, evaluations conducted in settings with higher quality of government may be associated with lower costs to evidence-based policy change. I explore these

patterns in figure 7, by considering responsiveness in election versus non election years; and high quality of government versus low-quality of government countries. I find no evidence of differential responsiveness across baseline political conditions.

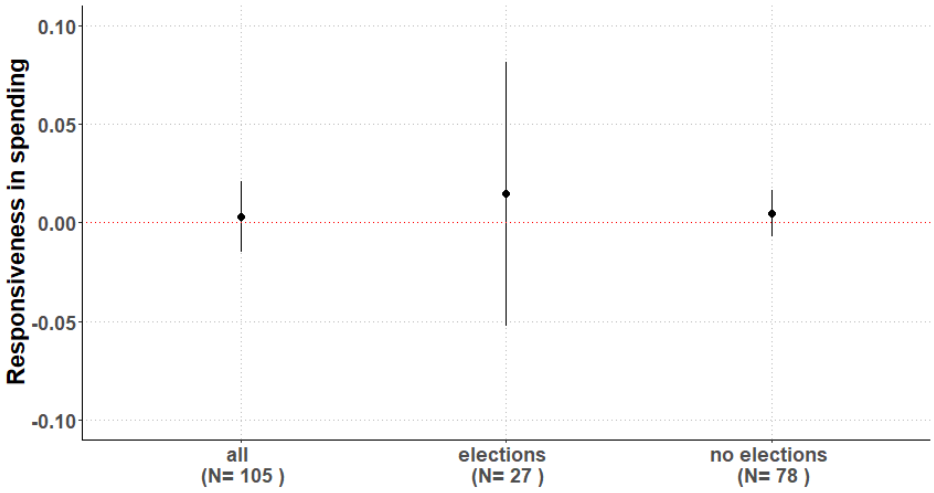


Figure 7: Linear relationship between TE and spending, by political conditions

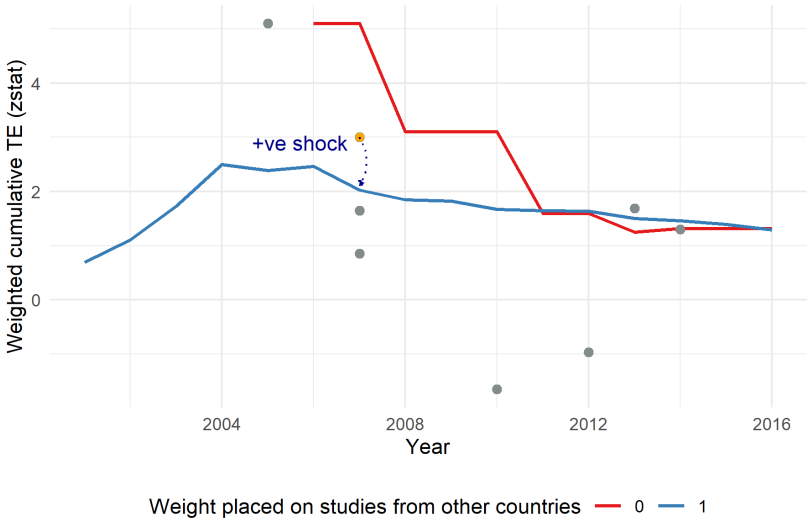
Notes: Linear relationship between evaluation outcomes and CCT spending, and 95% confidence intervals across subsets. *'All'* refers to the full sample; *'Elections'* refers to evaluations that are first published in an election year; *'No elections'* refers to evaluations that are first published in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

I find a robust zero association between spending and reported treatment effects across 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 4.2.

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 summarise each evaluation-level finding by the size of the shock, where I estimate an evidence-based prior belief as outlined in equation 6.

(a) $\lambda = 0$



(b) $\lambda = 1$

Figure 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, $\lambda = 1$).

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 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. As seen in panel a) where there is zero weight placed on research from other countries ($\lambda = 0$), the evaluation highlighted in orange is perceived as a negative shock (bad news), since the evaluation finding performs worse than 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.

I therefore estimate the relationship between evaluation surprises and changes in spending across different assumptions of λ, from 0 to 1. In figure 9 I plot the estimated $\hat{\beta}$ and 95% confidence intervals from a linear regression of equation 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 eexamine evidence for both of these channels, by considering the relationship between subsets of evaluation results that are more positive or negative, relative to the existing evidence-base (figure 10a, figure 10b). I find a consistent zero relationship for both positive and negative surprises.

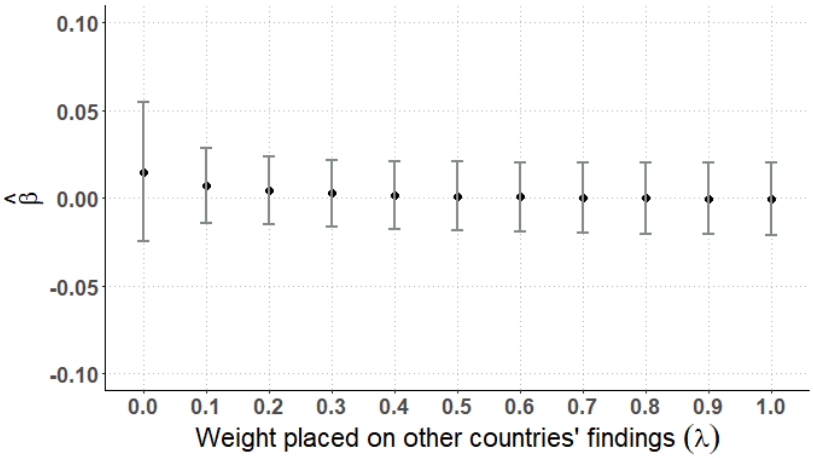


Figure 9: Relationship between quantified surprises and spending, across different assumptions on λ

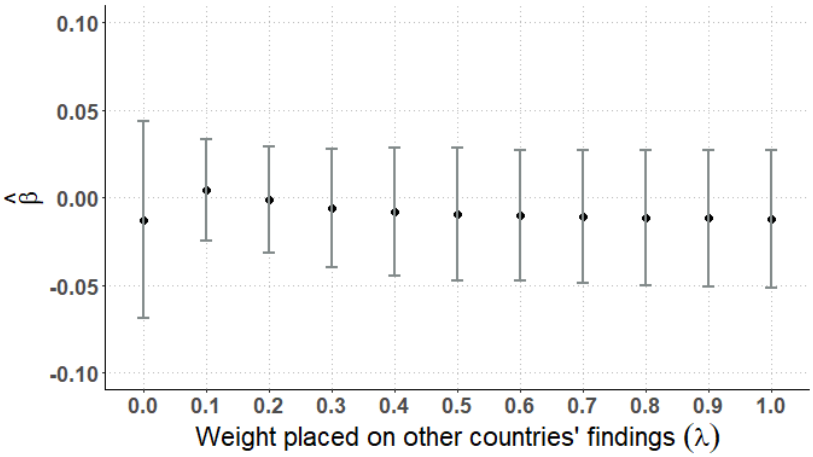
Notes: Estimated coefficient and 95% confidence intervals, for the linear relationship between quantified surprises and CCT spending, and across assumptions of λ. $\lambda = 0$: beliefs of zero external validity, i.e. zero weight is placed on research results from other countries; $\lambda = 1$ corresponds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

4.3 Framing of research results

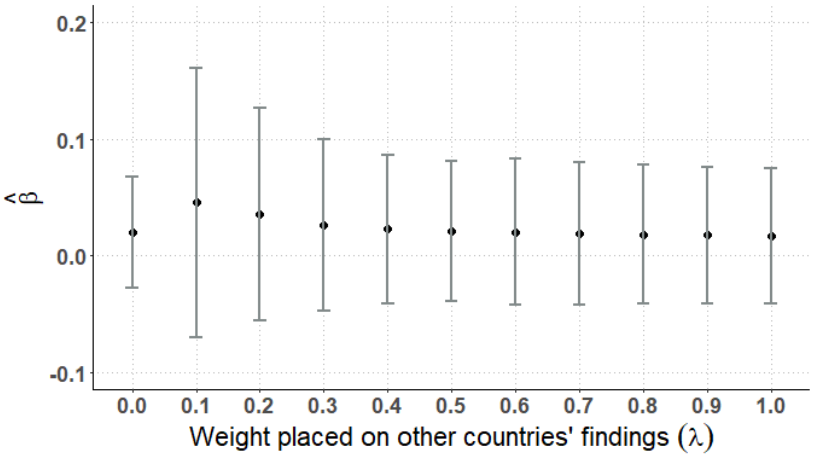
While I have thus far focused on treatment effects of headline findings, authors can also communicate the strength of evaluation outcomes through the language they use to describe the research results. I therefore move beyond aggregations of reported headline results, to consider policymaker responsiveness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

As outlined in section 3.1, I estimate the framing of research results by the sentiment score in the abstract (defined by equation 7). In general, the abstract sentiment score of evaluations tends to be positive, reflecting the idea that authors are included to use more positive than negative language to describe research findings. In figure 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 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



(a) Negative surprises only



(b) Positive surprises only

Figure 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.

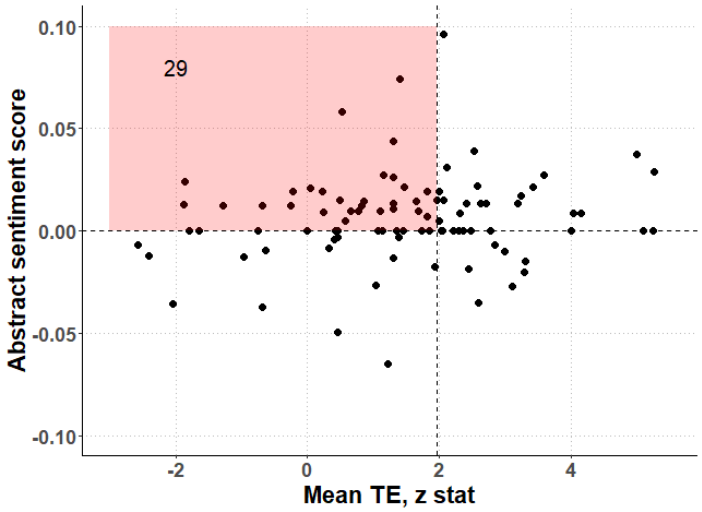


Figure 11: Mean Treatment effect (t-statistic) and the abstract sentiment score

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.

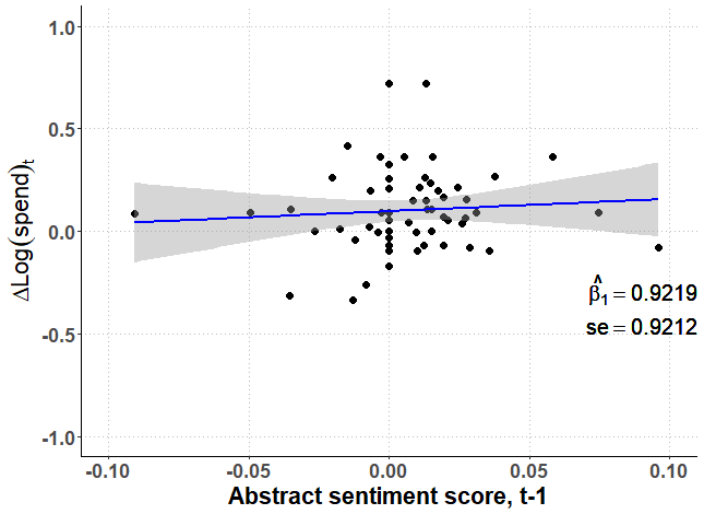


Figure 12: Abstract sentiment score and change in log spending

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.

evaluation results.

5 Cumulative evidence & spending

As seen in section 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 3.2, I estimate the posterior mean of each country's findings given the entire body of evidence.

In figure 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 [14]. 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).

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:

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

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

[14] 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.

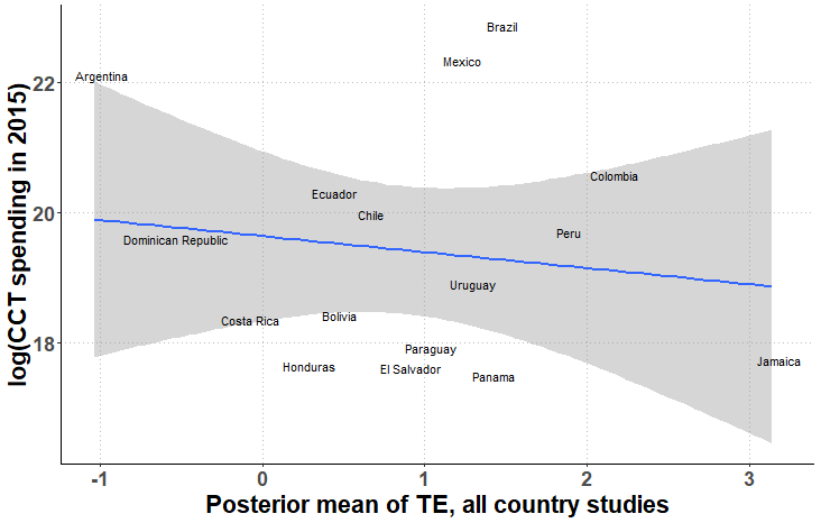


Figure 13: Bayesian posterior mean of aggregate results in 2015, and cash transfer spending

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

common mean – and hence, is suggestive of higher external validity.

Figure 14 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.

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 4) and cumulative country-level aggregations (section 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 constraints.

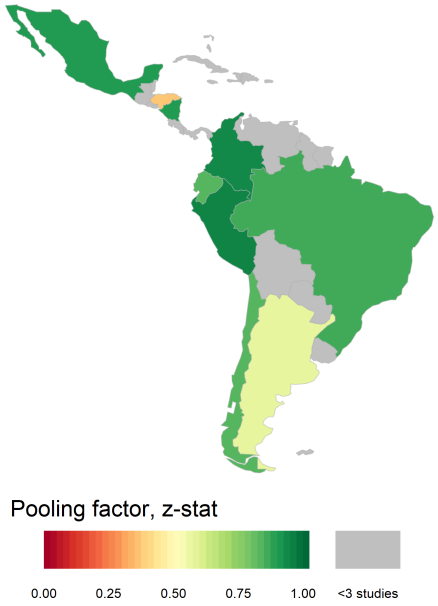


Figure 14: Estimated pooling factor of aggregate studies by country

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.

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 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]. The vast majority of these studies (135 studies out of 205) study the causal impact of Conditional Cash Transfers. Following CCTs, the most frequently evaluated programs are Unconditional Cash transfers (23 studies); and food transfers (12 transfers)[15]. Therefore, policymakers are not likely 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

[15]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.

expenditure. This suggests that the zero relationship is not driven by policymakers allocating more money to alternative social protection policies, in periods where evaluation outcomes on CCTs are higher.

Another challenge in interpreting the average zero relationship is the fact that the policymaker's objective function is unobserved. If the policymaker is not aiming to maximise the reduction in poverty (as in section 3), 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 causality. This would not be discernable from the data.[16]

Nonetheless, the zero average relationship can shed light on other objective functions that are common to the discussions around optimal 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 should be maximising on the cost-effectiveness of policies. If this is the case, an evidence-based policymaker maximising on cost-effectiveness would be reflected by a positive average relationship between evaluation results 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 CCT programs [Evans and Popova, 2016].

A final interpretation is related to the exogeneity of evaluations – or 'impact buying'. 32 evaluations in my sample are explicitly demanded by the 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 of evaluations, 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 driving the findings. First, I find that government demanded evaluations tend to be set up from the inception of the program. The evaluation of Progresa/Oportunidades

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

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 a 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 6).

7 Do features of evidence matter?

In section 3, I implicitly assume that all program evaluations are relevant to policymakers aiming to learn about the impact of their programs. If, however, obtaining and consuming evidence 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 a subset of evaluations that are more relevant for policy decisions. Importantly, the choice of policy relevant evaluations may further interact with political and practical constraints to policy change, K_{it}, as evaluations with certain characteristics may be associated with lower costs to evidence-use.

In this section, I consider heterogeneity 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 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 the study, policymakers may also place greater weight on studies that are peer-reviewed and published in top academic journals.

Even if policymakers do not learn disproportionately from more credible studies, there may be a higher association between treatment effects and spending for this subset of studies, if the constraints to evidence-use are lower for more credible evaluations. Randomised controlled trials are often referred to 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 15, I plot the association between the mean t-statistic and subsequent spending for subsets of studies, across different measures of evidence quality.[17] 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 observational 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[18]. 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. This means that evaluations that are internally valid but not broadly more generalizable are likely to less useful for policy decisions.

I measure the credibility of each study using the pooling factor from the Bayesian hierarchical model, given in equation 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.

[17]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.

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

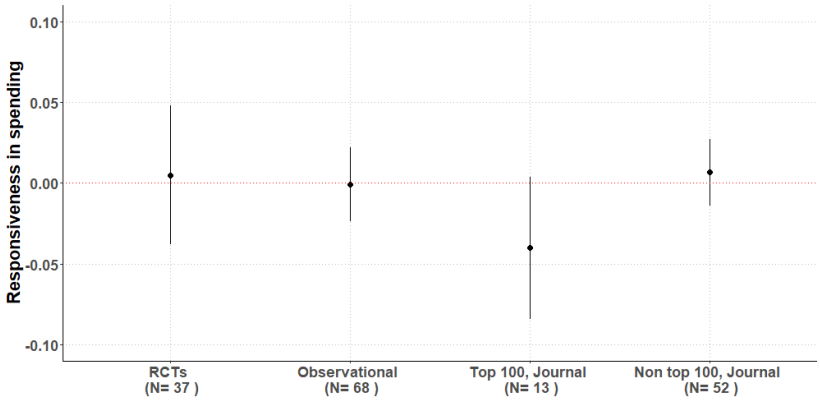


Figure 15: Responsiveness in spending to subsets of evidence, by credibility

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.

I also consider more direct measures of generalizability, 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 16, 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

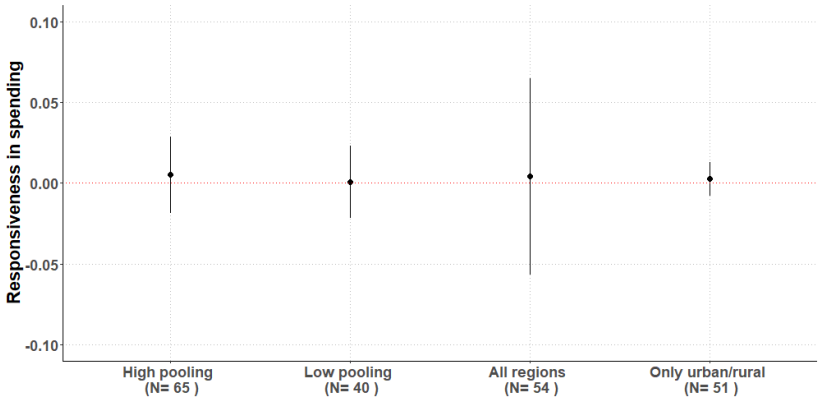


Figure 16: Responsiveness in spending to subsets of evidence, by generalizability

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.

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. In figure 17, I plot the association between spending and each of the outcome sub-categories and find a consistent zero relationship.

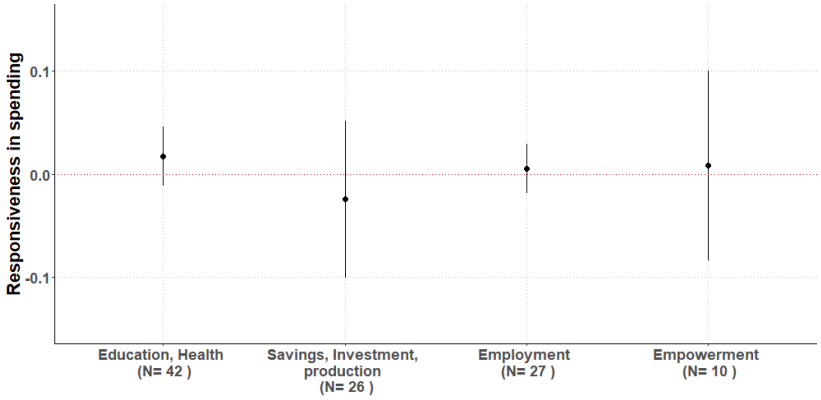


Figure 17: Responsiveness in spending to subsets of evidence, by outcome type

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

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.[19]

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 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.

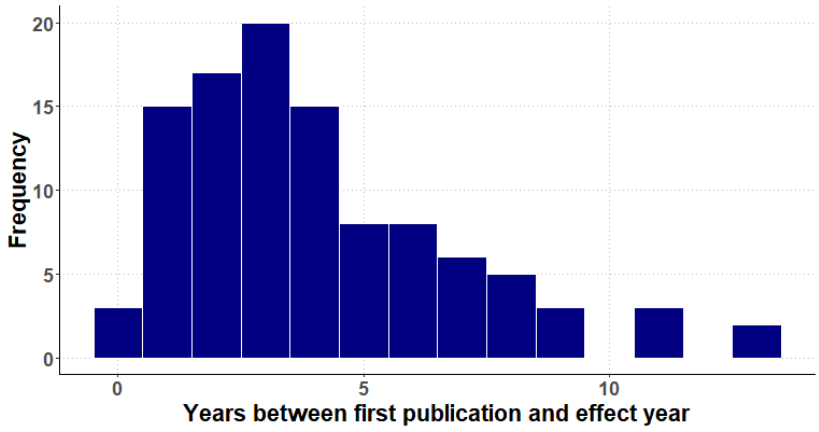


Figure 18: Timeliness of studies: distribution of number of years between the effect year, and the first year of publication

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

A longer lag between publication and the effect year is likely to correspond with lower actionability. This is because these studies are less likely to be embedded in the current policy environment. 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.

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

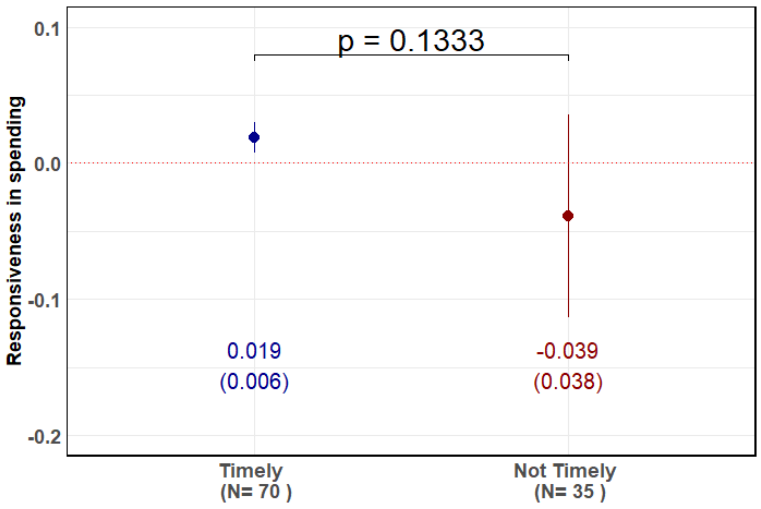


Figure 19: Responsiveness in spending to subsets of evidence, by timeliness

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.

As seen in figure 19, 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 result is not driven by the definition of timely studies. In figure 20 I show that the positive association persists for all studies that are published within the mean of 4 years before the endline evaluation.

	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 5: 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.

The importance of time-actionable results is driven by periods in which the political constraints to policy change is lower. In figure 21, I consider how the responsiveness in

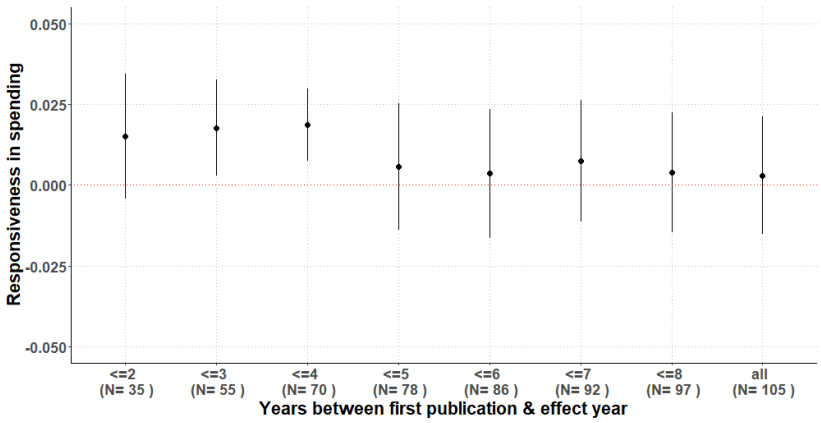


Figure 20: Responsiveness in spending by years between first publication and effect year

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

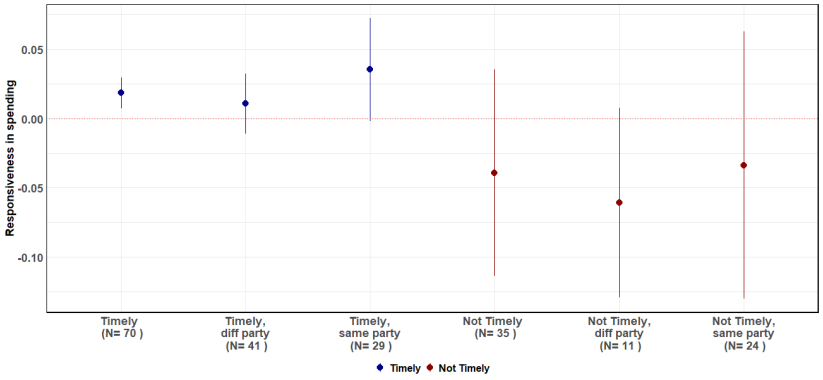
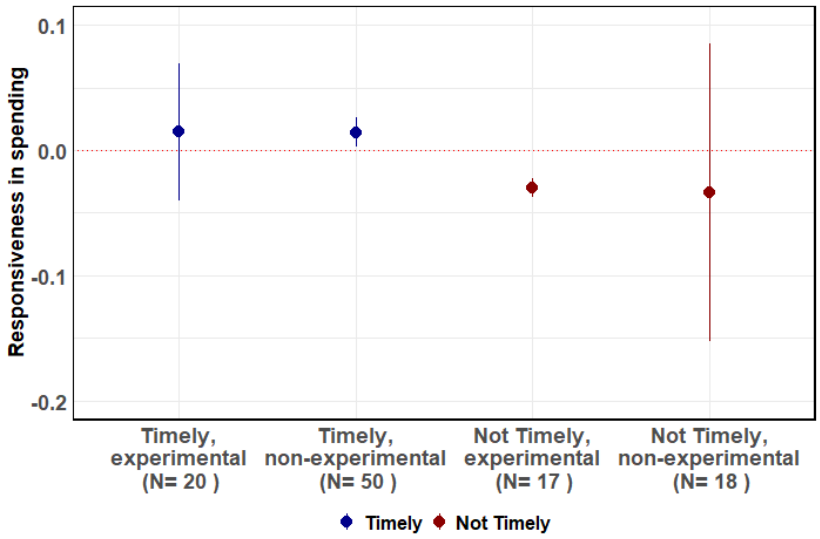


Figure 21: Responsiveness in spending to subsets of evidence, by timeliness and political party in power

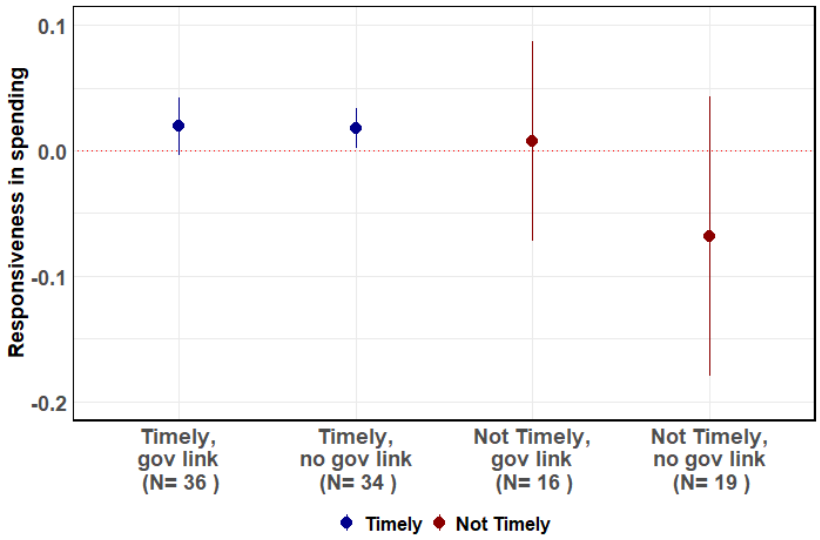
Notes: Linear relationship between evaluation outcomes and changes in spending, by timeliness and political party. *Timely*: evaluation is first published within four years of the effect year. *Sameparty*: the political party at the time of the first publication is the same as the party at the time of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

spending for timely evaluations interacts with changes in the political party in power. If the results of the evaluation can be attributed to the same political party as that in power at the date of publication, there may be greater political will to change policy in line with the evidence – and hence, lower political costs. 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 is higher when evaluations are timely, and when political constraints 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 5). In figure 22, 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.



(a) Responsiveness in spending by timeliness & methodology



(b) Responsiveness in spending by timeliness & government relationships

Figure 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.

8 Conclusion

Over the past two decades, there has been a vast increase in the number of program evaluations, particularly in lower and middle-income countries. In a world of limited resources, rigorous estimates of impact can help increase the efficiency of policy decisions and spending. However, my findings suggest that the potential benefits of these evaluations for policy spending have not been fully realised.

Across 128 program evaluations of Conditional Cash Transfers, I find a robust zero correlation between causal estimates of impact and subsequent policy spending. The only exception is when research results are time-actionable, and when political constraints are low. This suggests that the timeliness of publication 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 use of research in 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.

References

Hunt Allcott. Site selection bias in program evaluation. *The Quarterly Journal of Economics*, 130(3):1117–1165, 2015.

Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of economic perspectives*, 24(2):3–30, 2010.

Oriana Bandiera, Nidhi Parekh, Barbara Petrongolo, and Michelle Rao. Men are from mars, and women too: A bayesian meta-analysis of overconfidence experiments. *Economica*, 89:S38–S70, 2022.

Abhijit Banerjee, Esther Duflo, Nathanael Goldberg, Dean Karlan, Robert Osei, William Parienté, Jeremy Shapiro, Bram Thuysbaert, and Christopher Udry. A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236):1260799, 2015.

Sheheryar Banuri, Stefan Dercon, and Varun Gauri. Biased policy professionals. *The World Bank Economic Review*, 33(2):310–327, 2017.

Felipe Barrera-Osorio, Marianne Bertrand, Leigh L Linden, and Francisco Perez-Calle. Conditional cash transfers in education design features, peer and sibling effects evidence from a randomized experiment in colombia. Technical report, National Bureau of Economic Research, 2008.

Armando Barrientos and Juan Miguel Villa. Evaluating antipoverty transfer programmes in latin america and sub-saharan africa. better policies? better politics? *Journal of globalization and development*, 6(1):147–179, 2015.

Francesca Bastagli, Jessica Hagen-Zanker, Luke Harman, Valentina Barca, Georgina Sturge, Tanja Schmidt, and Luca Pellerano. Cash transfers: what does the evidence say. *A rigorous review of programme impact and the role of design and implementation features. London: ODI*, 1(7), 2016.

Alix Bonargent. Can research with policymakers change the world?, 2024. mimeo.

Francisco J Buera, Alexander Monge-Naranjo, and Giorgio E Primiceri. Learning the wealth of nations. *Econometrica*, 79(1):1–45, 2011.

Alberto Cavallo, Guillermo Cruces, and Ricardo Perez-Truglia. Inflation expectations, learning, and supermarket prices: Evidence from survey experiments. *American Economic Journal: Macroeconomics*, 9(3):1–35, 2017.

Simone Cecchini and Bernardo Atuesta. Conditional cash transfer programmes in latin america and the caribbean: Coverage and investment trends. Sep 2017. Available at SSRN: https://ssrn.com/abstract=3037640 or http://dx.doi.org/10.2139/ssrn.3037640.

Stefano DellaVigna, Woojin Kim, and Elizabeth Linos. Bottlenecks for evidence adoption. 2022.

Esther Duflo and Abhijit Banerjee. *Poor economics*, volume 619. PublicAffairs New York, NY, USA, 2011.

Esther Duflo and Michael Kremer. Use of randomization in the evaluation of development effectiveness[1]. *Evaluating development effectiveness*, 7:205, 2003.

Thad Dunning, Guy Grossman, Macartan Humphreys, Susan D Hyde, Craig McIntosh, and Gareth Nellis. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press, 2019.

Patrick Dylong and Fabian Koenings. Framing of economic news and policy support during a pandemic: Evidence from a survey experiment. *European Journal of Political Economy*, 76:102249, 2023.

David K Evans and Anna Popova. Cost-effectiveness analysis in development: Accounting for local costs and noisy impacts. *World Development*, 77:262–276, 2016.

Ariel Fiszbein and Norbert R Schady. *Conditional cash transfers: reducing present and future poverty*. World Bank Publications, 2009.

Alexander Frankel and Maximilian Kasy. Which findings should be published? *American Economic Journal: Microeconomics*, 14(1):1–38, 2022.

Sebastian Galiani and Patrick J. McEwan. The heterogeneous impact of conditional cash transfers. *Journal of Public Economics*, 103:85–96, 2013. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2013.04.004. URL https://www.sciencedirect.com/science/article/pii/S0047272713000789.

Andrew Gelman and Iain Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251, 2006.

Paul Gertler. Do conditional cash transfers improve child health? evidence from progresas control randomized experiment. *American economic review*, 94(2):336–341, 2004.

GiveDirectly. Cash evidence explorer. https://www.givedirectly.org/cash-evidence-explorer/, April 2023. Accessed: 2023-06-20.

Iasmin Goes and Stephen B Kaplan. Crude credit: The political economy of natural resource booms and sovereign debt management. *World Development*, 180:106645, 2024.

Aboozar Hadavand, Daniel S Hamermesh, and Wesley W Wilson. Publishing economics: How slow? why slow? is slow productive? fixing slow? Working Paper 29147, National Bureau of Economic Research, August 2021. URL http://www.nber.org/papers/w29147.

Johannes Haushofer, Paul Niehaus, Carlos Paramo, Edward Miguel, and Michael W Walker. Targeting impact versus deprivation. Technical report, National Bureau of Economic Research, 2022.

Jonas Hjort, Diana Moreira, Gautam Rao, and Juan Francisco Santini. How research affects policy: Experimental evidence from 2,150 brazilian municipalities. *American Economic Review*, 111(5):1442–1480, 2021.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

Independent Evaluation Group. *World Bank Group Impact Evaluations: Relevance and Effectiveness*. World Bank, Washington, DC, 2012. URL http://hdl.handle.net/10986/13100. License: CC BY 3.0 IGO.

International Initiative for Impact Evaluation, 3ie. 3ie Development Evidence Portal. URL https://developmentevidence.3ieimpact.org/. Accessed: 2024-11-05.

Stephen Brett Kaplan. Fighting past economic wars: Crisis and austerity in latin america. *Latin American Research Review*, 53(1):19–37, 2018.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Michael Kremer, Milan Thomas, Sasha Gallant, and Olga Rostapshova. Is development innovation a good investment? evidence on scaling and social returns from usaid's innovation fund. Technical report, Working Paper, 2021.

Ruth Levine and William Savedoff. Aid at the frontier: building knowledge collectively. *Journal of Development Effectiveness*, 7(3):275–289, 2015.

Rachael Meager. Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019.

Sultan Mehmood, Shaheen Naseer, and Daniel L Chen. Training policymakers in econometrics. Technical report, Technical report. Working Paper, 2021.

Nozomi Nakajima. Evidence-based decisions and education policymakers. 2021.

Laura B Rawlings and Gloria M Rubio. Evaluating the impact of conditional cash transfer programs. *The World Bank Research Observer*, 20(1):29–55, 2005.

Mark R Rosenzweig and Christopher Udry. External validity in a stochastic world: Evidence from low-income countries. *The Review of Economic Studies*, 87(1):343–381, 2020.

T Paul Schultz. School subsidies for the poor: evaluating the mexican progresa poverty program. *Journal of development Economics*, 74(1):199–250, 2004.

Mattie Toma and Elizabeth Bell. Understanding and increasing policymakers' sensitivity to program impact. *Journal of Public Economics*, 234:105096, 2024. ISSN 0047-2727. doi: https://doi.org/10.1016/j.jpubeco.2024.105096. URL https://www.sciencedirect.com/science/article/pii/S004727272400032X.

USAID. Strengthening evidence-based development: Five years of better evaluation practice at usaid 2011–2016, 2016.

Eva Vivalt and Aidan Coville. How do policymakers update their beliefs? 165:103121, 2023. ISSN 0304-3878. doi: https://doi.org/10.1016/j.jdeveco.2023.103121. URL https://www.sciencedirect.com/science/article/pii/S0304387823000767.

Shaoda Wang and David Y. Yang. Policy experimentations in china: the political economy of policy learning. 2021.

A Additional tables and figures

A.1 Individual evaluations & spending

Table A1: 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 A2: 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 A3: 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 A4: 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 A5: 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.

A.2 Cumulative results & spending

Table A6: 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.

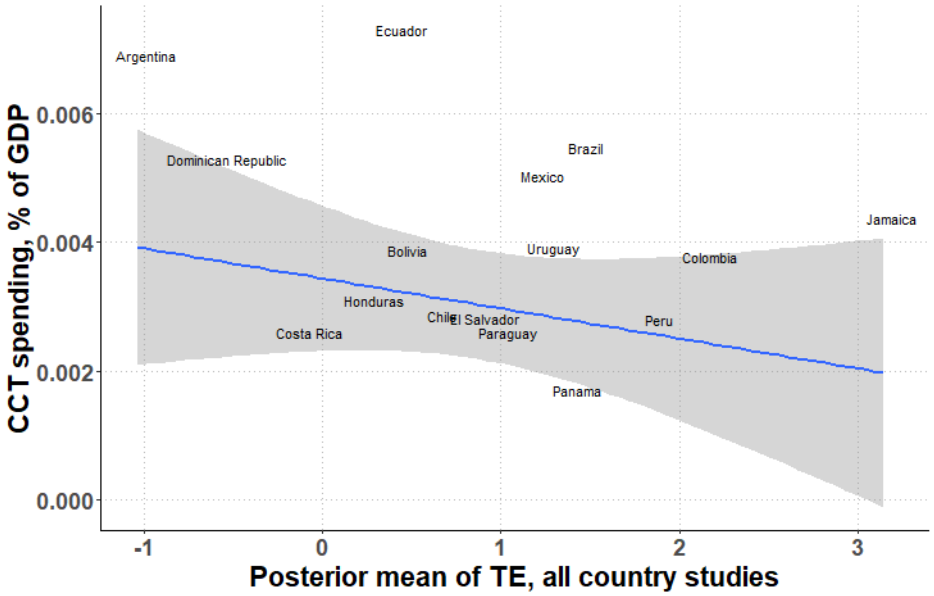


Figure A1: Posterior mean of treatment effects and spending, as percentage of GDP

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.

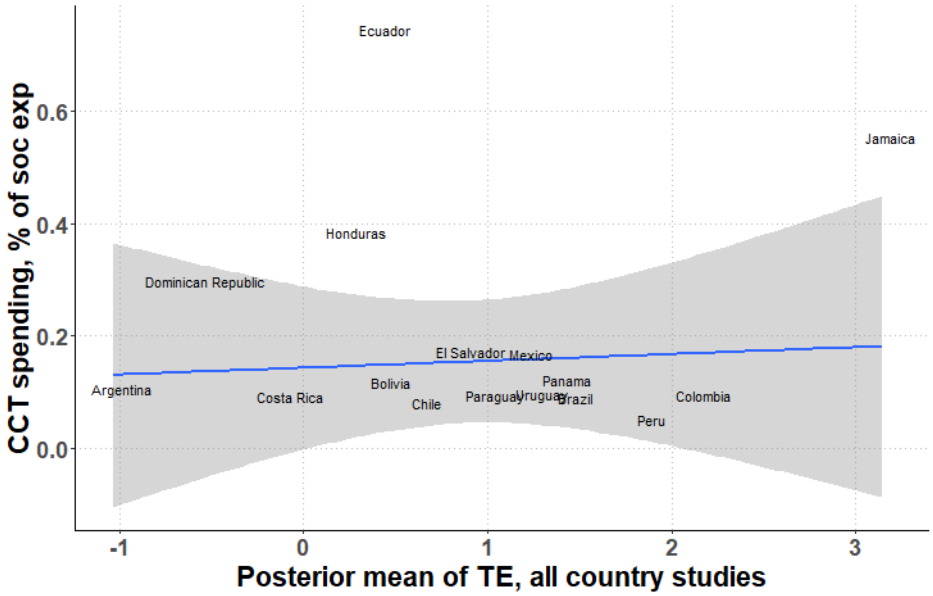


Figure A2: Posterior mean of treatment effects and spending, as percentage of social protection

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.

A.3 Features of evidence

Table A7: Relationship between measures of 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. *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 A8: Relationship between measures of 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. *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 A9: Relationship between measures of 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 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 A10: Relationship between measures of 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. *Timely*: evaluation is first published within four years of the effect year. Standard errors are clustered at the country level.

Table A11: Relationship between measures of 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. *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 A12: Relationship between measures of 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. *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.

B Additional details on data

B.1 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 A13: 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

B.2 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.

C Additional results

C.1 Policymaker background and spending

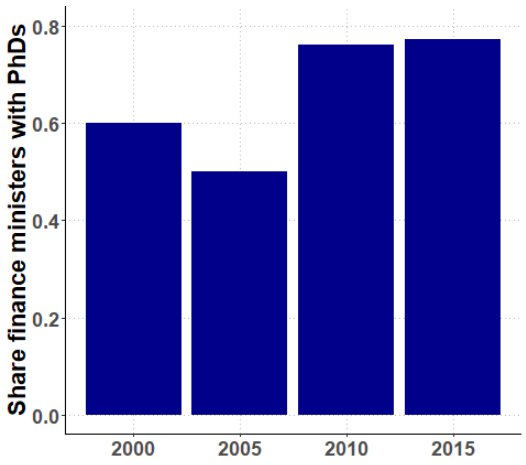


Figure A3: Proportion of finance ministers with PhDs in Latin America and the Caribbean, by year

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].

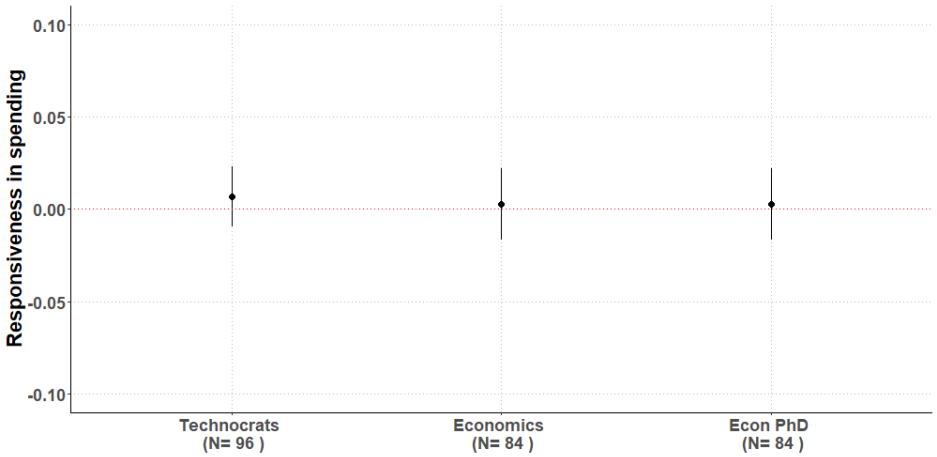


Figure A4: Relationship between mean t-stat and spending, by finance minister training

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.

C.2 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. In the figure below, I consider relationship between evaluation outcomes and subsequent changes in spending for evaluations that are more timely, where I assume that the first date at which policymakers may be aware of the evidence is the effect year.

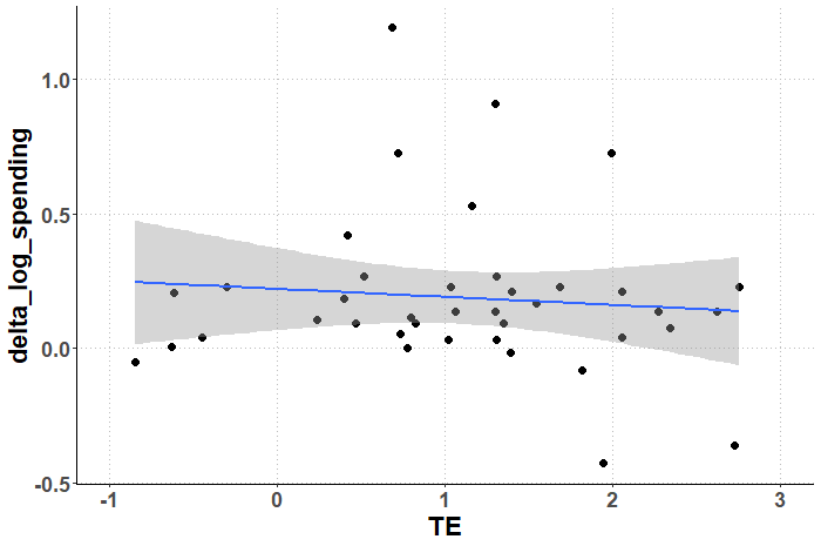


Figure A5: Relationship between mean t-stat and subsequent spending, matched by the endline year of the evaluation

Notes: Linear relationship between mean t-stat and changes in spending, one year after the effect year of evaluation. Consider only the subset of evaluations that are published within three years of the effect year.

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

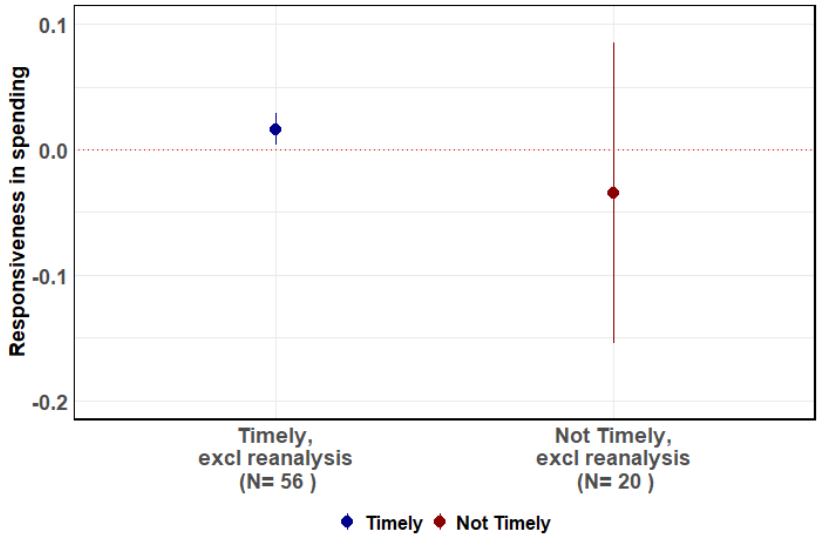


Figure A6: Relationship between mean tstat and subsequent spending, excluding studies that use experimental data from prior RCTs

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.