Bayesian Impact Evaluation with Informative Priors: An Application to a Colombian Management and Export Improvement Program *

Leonardo Iacovone (World Bank) David McKenzie (World Bank)
Rachael Meager (LSE)

December 15, 2022

Abstract

Policymakers often test expensive new programs on relatively small samples. Formally incorporating informative Bayesian priors into impact evaluation offers the promise to learn more from these experiments. We evaluate a Colombian program for 200 firms which aimed to increase exporting. Priors were elicited from academics, policymakers, and firms. Contrary to these priors, frequentist estimation can not reject null effects in 2019, and finds some negative impacts in 2020. For binary outcomes like whether firms export, frequentist estimates are relatively precise, and Bayesian credible posterior intervals update to overlap almost completely with standard confidence intervals. For outcomes like increasing export variety, where the priors align with the data, the value of these priors is seen in posterior intervals that are considerably narrower than frequentist confidence intervals. Finally, for noisy outcomes like export value, posterior intervals show almost no updating from the priors, highlighting how uninformative the data are about such outcomes.

JEL Classification: C11; C93; F14; O12; O14

Keywords: Bayesian Impact Evaluation; Prior Elicitation; Randomized Experiment; Management

*We thank: Darío Rodríguez Pérez for his collaboration on the initial stages of this project; staff at the Colombian government agencies PTP and DNP; the academic and policy experts who provided priors for this project; Andrew Gelman for feedback on our pre-analysis plan; seminar audiences at BREAD, Chicago, Dartmouth, DeNeB, Fudan, Maryland, St Andrews, and Yale; and Juan Sebastián Leiva, Daniela López, Sofía Jaramillo, Marta Carnelli, Maria Juliana Otalora and Kyle Holloway of IPA Colombia. We gratefully acknowledge funding from the Knowledge for Change Program and the World Bank's CIIP Trust Fund. This study was pre-registered in the AEA Social Science Registry on June 26, 2018 (AEARCTR-0003109) and was accepted as a stage 1 registered report at the Journal of Development Economics.

1 Introduction

Governments and researchers often test new policies by experimenting on a relatively small number of units. This is particularly common with government programs intended to help small and medium enterprises, as these interventions are expensive. For example, Bloom et al. (2013) piloted a management improvement program in India with 28 plants tracked for two years; Higuchi et al. (2017) a program with treatment groups ranging from 26 to 133 firms in Vietnam and tracked for five years; Bruhn et al. (2018) a program with 150 treated firms in Mexico tracked for four to five years; Iacovone et al. (2022) a program with a 159 firms in Colombia tracked for four years; and Custódio et al. (2020) a program with 93 firms in Mozambique, tracked for a year. However, these small samples can result in analyses with limited statistical power, and the estimates of the program's impacts may be imprecise, with standard frequentist hypothesis testing unable to reject the null of no impact at conventional significance levels even if the point estimates are positive and of a magnitude that would pass a cost-effectiveness test.

But these policy interventions are not designed in a void. Policymakers design interventions based on their past experiences with different policies in the country and their knowledge of the context and constraints facing beneficiaries. Academics bring knowledge from the existing literature, economic theory, and their own experiences. Participants - such as firms - who apply to such programs do so based on expectations about their likely effects. These beliefs typically involve considerable uncertainty (since if they knew for sure the program would work as intended, a pilot would not be necessary). But nevertheless, they contain some information, which standard impact evaluations using frequentist estimation completely ignores.

Bayesian analysis theoretically offers a principled way of incorporating this prior knowledge into impact evaluation. Imbens and Rubin (2015) provide the foundational treatment of a Bayesian model-based econometric approach for inference in simple randomized trials. Yet like many textbook approaches to Bayesian analysis, including Gelman et al. (1995), they tend to focus on weakly-informative priors in large-sample applications, rather than relatively informative priors in small sample applications. Conducting informative Bayesian impact evaluation in a real applied setting raises a number of practical challenges around the choice of estimand, how informative priors should be obtained, how to deal with large numbers of control variables such as randomization strata fixed effects, and related modelling challenges.

This paper provides a demonstration of how to apply Bayesian impact evaluation in practice to a real stakes field experiment. We consider an evaluation of a program in Colombia that was designed to help increase exports by improving management practices. This tackles a key policy question - how to diversify exports and increase firms'

productivity - in the context of a multi-million dollar experiment on a sample of 200 firms. We elicit informative nonparametric priors from academic experts, Colombian policymakers, and from the firms participating in the program, as well as using literature-informed priors. Our priors are elicited for the intention-to-treat (ITT) effects, we then aggregate them within category (e.g. policymaker), and fit distributions to these priors. Our frequentist estimation employs five lags of the outcome variable and stratified randomization with over fifty strata, following standard practices typically employed in the literature to boost statistical power with small samples. Our Bayesian estimation then regularizes these strata fixed effects, and combines the informative priors with a regression model using the experimental data in order to obtain credible posterior intervals of treatment effects. We compare the results to those using non-informative (diffuse) priors to show how much of the results are coming from the Bayesian methodology versus the information contained in the priors.

Take-up for the program was high, with 83 percent of firms offered treatment receiving consulting services, with a median of 195 hours consulting received. However, we find that while this consulting improved general management practices in the firm (particularly lean manufacturing operations practices and customer-related commercial practices), it led to no change in export-specific practices that are specifically geared towards improving export outcomes. Our frequentist estimates find small and statistically insignificant impacts of the program on exporting in 2019 (the first year after the intervention started), and statistically significant negative impacts on some export outcomes in 2020 (the second year after intervention, and one in which export markets were also disrupted by the COVID-19 pandemic).

These estimates differ from the positive impacts of the program anticipated by academics, policymakers, participating firms, and the existing literature. Our Bayesian impact evaluation results show how much these priors should be updated in light of the data from this pilot experiment. For binary outcomes such as the extensive margin of whether firms export at all, our frequentist estimates are relatively precise, and the Bayesian credible posterior intervals overlap almost completely with traditional confidence intervals. That is, even with informative priors, the signal in the data is strong enough to completely update priors on these outcomes. For outcomes like increasing export variety, where the priors are in line with the experimental results, the value of incorporating these priors is seen in posterior intervals that are considerably narrower than the frequentist confidence intervals, providing more precision about the likely effect sizes. In contrast, our sample of firms is very heterogeneous, differing a lot in size and export performance, making it hard to detect changes in skewed continuous outcomes. As a result, our frequentist estimates of the impacts of the program on the value and productivity of exports are very imprecise, with wide confidence intervals. In this case, the Bayesian posterior intervals

show almost no updating from the priors, highlighting how uninformative the data are about such outcomes. Finally, we construct an overall index of export performance as a sum of standardized z-scores of our pre-specified export outcomes. Our Bayesian credible intervals lie between the prior intervals and confidence interval, showing partial updating. The results show the program was not as successful as anticipated, but also temper the conclusion of it perversely worsening export outcomes.

The Bayesian approach to impact evaluation also offers several natural extensions to the analysis above which allow us to go beyond estimating a single point treatment effect and testing whether it is zero or not. The most straightforward additional analysis is to calculate the posterior probability that the program had an impact large enough to pass a cost-benefit analysis.[1] In our setting, given the data, we find that the chance that the program has an average effect large enough to justify the government's spending is typically much less than 10%. However, it is possible that this average effect is composed of heterogeneous "individual treatment effects" (ITEs) for the firms we consider; within the Bayesian framework it is possible to directly estimate the distribution of ITEs by modelling the potential outcomes, as proposed in Imbens and Rubin (2015). This analysis is natural in the Bayesian set-up and challenging in the frequentist approach because it requires information about the correlation between the potential outcomes, which the data can never speak to (due to the fundamental problem of causal inference). We use elicited priors on this correlation from academics and policymakers to perform this analysis in our sample and find limited potential for heterogeneous effects of the program.

There are several possible explanations for this somewhat negative result, mostly relating to program implementation. A first-order issue is the type of consulting advice provided: while the government initially focused on export outcomes, over time the program goals appeared to shift towards general management practices and firm productivity overall. Our qualitative discussions with consultants (Appendix B) suggested that in some cases they suggested firms were not at the standard needed to export and should focus instead on the domestic market; this may explain the negative effect on exports in 2020.[2] Some relevant additional issues include low program intensity relative to similar SME interventions studied in the literature, and the fact that our sample of firms was more heterogeneous in both size and industry, making it potentially more challenging for consultants to offer relevant advice to all firms. We explore these narratives and perform robustness checks to our main results by excluding strata with exceptional export performance at baseline in appendix E. Taken together, our results suggest that these features of program implementation may be important to program success.

[1] While this is sometimes done in frequentist analysis, there is little justification for doing so, as the frequentist uncertainty refers to the distribution of the *estimator* across multiple samples, not the distribution of uncertainty about the estimand itself.

[2] While we find some suggestive evidence that other firm outcomes may have improved, the evidence is quite weak; see Appendix E.

This paper serves to demonstrate through an application to a real world policy experiment how Bayesian impact evaluation can be carried out in practice with informative priors. Our hope is that it serves a similar role to Casey et al. (2012), who demonstrated the practical use of pre-analysis plans through a field experiment. Just as with pre-analysis plans, we believe the value of this Bayesian approach is likely to be highest for long-term field experiments that are expensive and time-consuming to replicate. In contrast, if replications are cheap and quick, as is the case in many lab experiments, then just as Coffman and Niederle (2015) argue against pre-analysis plans, one may be able to cheaply increase sample size to the point where posterior credible intervals coincide with traditional confidence intervals, even with informative priors, reducing the value of our approach. The same logic may apply to large sample A/B testing carried out on tech platforms. But when important policy decisions rest on the results of field experiments with limited sample sizes, investing in collecting informative priors and bringing this knowledge into evaluations through our approach seems beneficial.

This paper contributes to two main areas of the literature. The first is a growing literature on collecting and using priors and predictions about interventions. To date this has largely involved getting academics, policymakers, or program participants to make predictions about the likely effects of a treatment, and then measuring the accuracy of these predictions (e.g. Groh et al. (2016); Hirschleifer et al. (2016); Dellavigna and Pope (2018); McKenzie (2018); Dellavigna et al. (2019)). But in economics these predictions have typically not been elicited as full nonparametric distributions, and have not been formally incorporated into the analysis. Andrews and Shapiro (2021) study the question of how to best communicate results if audiences come with different priors, and suggest that there may be cases where analysts should censor the estimates they report as a result. Abadie (2020) notes that the information content in non-significant results will depend on the priors decision-makers had about the null hypothesis being rejected. Our approach explicitly elicits these priors, and so provides a way to communicate how much these priors should be updated given the data. Finally, while Bayesian analysis has not yet been used for analyzing individual impact evaluations, Bayesian meta-analysis approaches have been used to aggregate evidence from different studies (e.g. Meager (2019); Vivalt (2020)).

Second, the paper contributes to a literature on improving management and export performance in firms. Several recent experiments have found management consulting can improve management practices and firm performance (Bloom et al. (2013); Bruhn et al. (2018); Iacovone et al. (2022)). However, these existing studies have not focused on export-oriented firms with exporting as a main outcome. Bloom et al. (2021) use data on American and Chinese firms, and find that better managed firms are more likely to export, sell more products to more destinations, and have higher export revenues. Despite this relationship, our results suggest that purely improving management practices may

not be enough by itself to spur export activity without a more explicit focus on practices related to exporting. To date there have been relatively few rigorous evaluations of these policies to spur exports in developing countries. Several studies use ex-post evaluations with non-experimental methods (e.g. Girma et al. (2009) on production subsidies for Chinese firms, and Cadot et al. (2015) on matching grants given to Tunisian firms to implement export business plans), but concerns remain about self-selection of firms into these programs. Atkin et al. (2017), carried out a demand-side intervention with small firms (average size of one employee), providing Egyptian rug-manufacturers with initial orders and links to foreign buyers, and find exporting increases firm productivity. Two other experiments do not find statistically significant impacts of much lighter information interventions: Breinlich et al. (2017) consider the impact of sending brochures from the export promotion agency to SMEs in the United Kingdom, and Kim et al. (2018) find that one-day informational seminars had no impact on exporting for textile firms in Vietnam. Even though our intervention is much more intensive, it also shows how hard it can be for supply-driven approaches to increase export performance.

The remainder of this paper is organized as follows. Section 2 describes the experimental context and intervention. Section 3 outlines how informative priors were elicited from academics, policymakers and firms, and compares them to literature-based and weakly-informative priors. Section 4 describes the methodological approach for both frequentist and Bayesian impact evaluation. Section 5 provides all results of the main analysis, with additional Bayesian decision analysis in subsection 5.4 and a discussion of why the program impacts were not as large as expected in subsection 5.5. Section 6 concludes with potential applications of the informative Bayesian approach in future research.

2 Context, Experiment and Intervention

2.1 Background and Program Launch

Colombia faces several important policy challenges which this program aimed to help address. The first is the desire to diversify and expand its export base. Colombia is currently highly dependent on commodities such as crude petroleum, coal, coffee, flowers, and gold, which make up more than 80 percent of its merchandise exports, resulting in a policy interest in broadening the range of firms and sectors that engage in exporting. Secondly, one key barrier to firms being competitive on the global market is low productivity. Labor productivity in Colombia is low, with it taking around four Colombian workers to produce what one worker does in the United States (Londoño, 2017). As a result, improving productivity is a priority for government policy, and helping address this productivity goal may also help firms become more competitive in global markets.

A key factor that helps determine both firm productivity, and the ability of firms to export, is the management practices used by firms. Bloom et al. (2016) estimate that differences in management can account for 30 percent of cross-country productivity differences. Better management is also strongly associated with export performance, with Bloom et al. (2021) finding that better managed firms are more likely to export, sell more products to more destinations, and have higher export revenues. However, Colombian firms have low levels of management practices according to global surveys like the World Management Survey, with levels similar to countries like India and Kenya that have much lower per-capita incomes. A prior pilot experiment in Colombia (Iacovone et al., 2022) had worked with a single industry, the autoparts sector, and found that individual and group-based consulting had improved management practices, but was not focused on exporting or export-oriented firms.

The Colombia Productivity and Export Improvement Program (PEIP)[3] was designed by the Programa de Transformación Productiva (Program of Productive Transformation) (PTP) with the explicit objective of "improving the productivity and export capacity of the participating firms". It aimed to achieve these goals through providing consulting services to firms to improve their management practices (described in detail below). The program was launched in November 2017. To be eligible for the program, firms had to have existed for at least two years, be formally registered, belong to one of thirteen selected sectors (transport manufacturing, construction, textiles, fruits and fruit products, specialty coffees and coffee products, beef, aquaculture, cocoa products, processed food, cosmetics, pharmaceuticals, plastics and paint products, and basic chemicals), provide financial statements, and complete an online application process. Along with the general call for firms to apply, PTP specifically reached out to a list of firms that had exported at least once in the past five years to see if they were interested in applying. The closing date for applications was March 23, 2018 (see timeline in Appendix A).

2.2 Random Assignment and Firm Characteristics

A total of 200 firms met the eligibility conditions for the program after applying. These 200 firms were then randomly assigned to two groups of 100 each on April 11, 2018. The application form data were used to stratify firms by size (small, medium, or large), and whether or not the firms had exported at all in the last 3 years. An additional two strata were added: one stratum of 19 export outlier firms (defined in terms of having export values, the number of destinations exported to, or the number of different products exported above the 95th percentile in the self-reported export data on the application form), and one stratum of 1 firm that was missing firm size information. We then formed

[3]This is a pseudonym, which we use at the request of PTP.

an index of the proportion of 11 exporting management practices that firms were using. Within each of the eight strata, we then ranked firms by this export practices index, and formed quadruplets, randomly selecting two firms from each quadruplet to be assigned to control, and two firms to treatment. In total this gives us 54 strata defined by these export practice quadruplets inside the eight original strata.

Table 1 provides summary statistics for the firms in the program, providing means by treatment status, along with the standard deviation and percentiles of the distribution. There are three features of the sample that we wish to emphasize. The first is that the sample consists of firms that were already exporting or that were interested in doing so. At the time of applying, 58% had exported in the last three years, and 50% had exported in the last year (2017). Conditional on exporting, the median firm exports $170,000, exporting 3 different products (measured at the 6 digit level), to 2 countries, and a total of 5 distinct product-country combinations. Among the firms that were not exporting, 95 percent said they were interested in starting to export, and among those already exporting, 99 percent said they were interested in expanding exports. When asked to describe how they see the strategic role of exports in their business, 81 percent of those exporting saw it as a key part of their firm growth strategy, as opposed to a means of risk diversification, a way of selling excess products that failed to sell locally, or an occasional response to requests from abroad.

Second, firms had room to improve their export and general management practices. On average, firms are doing 36 percent of the basic export practices measured on the application form, and 44 percent of general management practices. Third, the firms are very heterogeneous, which makes it harder to offer a standardized program, and also more difficult to detect treatment impacts. This heterogeneity is evident in firm size (47 percent small, 45 percent medium, 9 percent large), with firms having a mean of 73 and median of 42 employees, but ranging from 2 to 750 workers. The sales of a firm at the 90th percentile (25,675 million pesos or US$8.6 million) are more than five times those of a firm at the median (4,596 million pesos or US$1.5 million), and 36 times those at the 10th percentile (700 million pesos or US$235,000). This heterogeneity also shows up across sector, with the most common sectors being textiles (18%), construction (16%), transportation equipment (14%), plastics and paint products (13%), and processed food (10%). Specific examples include a clothing factory specializing in business and school uniforms; a manufacturer of window frames out of aluminum; a cosmetics factory making shampoo and nail polish; a paint manufacturer; and a company making dried tropical fruits. Almost half of the firms are located in the Cundinamarca region that includes the capital city, Bogota, with another quarter in the two regions around Cali and Medellin.

2.3 Details of the Intervention and Program Implementation

The program began with both the treatment and control groups receiving a diagnostic analysis in April and May 2018. This consisted of a consultant assessing the firm in five areas: quality (getting products to the standard needed for international markets), productivity (methods to reduce production costs), labor productivity (a focus on managing workers to make them more productive), commercial strategy (with a focus on sales strategy and accessing export markets), and energy efficiency (to reduce energy costs of production). The diagnostic was free for firms, but involved about 12 hours of consultant time, at an approximate cost to the program of US$625, and concluded with an automatized summary report which compared their relative performance across the five areas, along with providing general (not customized) high-level advice for areas for improvement in each area.

The treatment group then received a consulting intervention, beginning in June/July 2018 and lasting for most firms through to July 2019. This consisted of 190 hours of in-person technical assistance: 30 hours directed towards general commercial strategy, and 160 hours towards the two of the other four areas. The diagnostic guided this choice, but firms were free to choose which two areas they preferred. Each of the five areas was contracted separately to a different Colombian consulting firm that specialized in that particular area, and who would then send consultants to work with the firm. The commercial strategy consulting focused on identifying a star product and determining which markets the company should devote its sales efforts (both domestically and internationally). The operational productivity consulting involved implementing lean manufacturing tools like value-stream mapping with the goal of standardizing processes, reducing bottlenecks, and improving production efficiency. The labor productivity consulting focused on retaining and improving worker morale, through methods such as worker recognition programs and feedback sessions. The quality consulting focused on improving quality standards to the level needed to meet technical barriers to enter overseas markets. The energy consulting looked for opportunities to lower energy costs through improvements, such as through using LED lighting. Appendix B provides more details of the consulting intervention along with our qualitative observations on each component. This consulting treatment is estimated to have a market value of 40 million Colombian Pesos (approximately US$13,800). Small firms selected for the program had to pay 3 million COP ($1,035) and medium and large firms 6 million COP ($2,070), which could be paid in multiple installments.

The initial design of the program by PTP included a theory of change in which the program would first improve management practices, standardize processes, and reduce energy costs. This was then expected to lead to firms having more products available that meet export quality requirements, resulting in an increase in both the amount and variety of exporting. The most direct route to improving exports would involve improving

export-specific management practices, such as designing quality and production targets specifically for external markets, and directing commercial and marketing efforts towards acquiring customers and distributors in overseas markets. The more indirect route is to focus instead on general management practices intended to help firms become more productive, which will help increase their competitiveness and make them better able to compete in foreign markets. A change in management focus within the program agency (PTP) and in broader government strategy meant that some of the emphasis of the program shifted more towards this general productivity focus and indirect route to exporting, and less on direct export practices as implementation was taking place.

3 Eliciting Priors: What, How, and From Whom?

Using informative priors in Bayesian impact evaluation requires finding a way to bring in outside knowledge, that is external to the data collected for frequentist analysis. One source of outside knowledge could be the existing literature, which is the approach advocated by Gelman et al. (1995) for Bayesian data analysis. However, in many large-scale field experiments, the existing literature may have few, if any, studies with a similar-enough intervention and context, raising concerns about the external validity of existing evidence. Moreover, the evidence in the existing literature may differ from the knowledge and beliefs of the key decision-makers involved in the policy being evaluated. An alternative approach is to instead elicit priors from key experts and decision-makers.

In either case, researchers will need to decide what priors they wish to obtain. Eliciting priors then requires also deciding how to elicit these priors and put them in a form suitable for use in analysis, and from whom to elicit this information. We discuss our choices, and considerations for other researchers in obtaining priors.

3.1 What priors should be elicited?

Our pre-analysis plan and registered report defined eight primary outcomes, intended to measure whether the program caused more firms to export, diversified the range of products exported and destinations exported to, and improved the export performance of participating firms. In most experiments, there will be some non-compliance, and so researchers will need to decide if they want to obtain priors on the intention-to-treat (ITT) effect, or on the local average treatment effect (LATE) for their primary outcomes.

In some respects and contexts, the LATE may be a more natural object for many policymakers and program participants to think about, since it captures the impact of a

program who actually take it up when offered.[4] However, if respondents differ in their beliefs about the take-up rate (or on who the compliers may be), then the priors obtained from different respondents may not be directly comparable to one another, and if their beliefs about take-up are inaccurate, may also correspond to a different parameter than the LATE estimated for the study sample. Aggregating LATEs over different respondents would require eliciting bivariate distributions over both LATEs and take-up rates, which would be complicated for most respondents to manage.[5]

Instead, we believe that the ITT is easier to elicit from respondents, and, as we discuss below, can be easily aggregated across respondents to take advantage of the wisdom of crowds.[6] These priors capture the joint distribution of beliefs over take-up rates, and of impacts for those who do and do not take-up the intervention. In eliciting beliefs, we clearly explain to respondents that we are interested in the ITT, and explain this as the difference in outcomes for the full group of 100 firms offered treatment compared to the full group of 100 firms in the control group. We walk them through an example to make sure they understand that this also includes impacts for those who do not take up the program (see Appendix D for language used).

In addition to eliciting the ITT for our eight primary export outcomes, we also elicited priors over two additional parameters. The first was the take-up rate. We do not use this in estimation at all, but use it as a check to see whether the actual take-up rate was at least as higher as anticipated by respondents when they gave their priors over the ITTs. Finally, one of our two approaches to Bayesian estimation requires fitting likelihood models that require fitting the joint distribution of potential outcomes with and without treatment. This joint distribution depends on the unobserved correlation between a firm's export performance with the program and without the program. We therefore also asked our academic experts for their priors on this correlation.

3.2 How should priors be elicited?

Economists have increasingly shown interest in predicting the results of experiments, although many current applications just ask for point estimates and potentially some metric of uncertainty (e.g. Dellavigna et al. (2020), Dellavigna et al. (2019)). Instead, for

[4]This assumes that the exclusion restriction holds, where there is no effect of the program for those who are offered it but do not take it up. This seems plausible in our context, but may not hold in some other settings in which researchers are interested in using Bayesian impact evaluation.

[5]Another possibility, when the exclusion restriction is believed to hold, would be to wait and elicit the LATE after take-up is known, presenting respondents with both the take-up rate and a table of observable characteristics of compliers.

[6]We also note that eliciting the ITT for a linear model will be easier than eliciting priors for treatment effects estimated using nonlinear models. For example, priors for the marginal effects from probit estimation will depend on both what individuals expect the control mean to be, as well as what they believe the marginal effect will be when calculated at that control mean.

our priors we require eliciting probabilistic expectations about the full distribution of the ITTs. Manski (2004) pioneered the collection of probabilistic expectations, and Delavande et al. (2010) discuss different approaches for implementing it in a developing country setting. Prior elicitation is regularly done for clinical trials in medicine, pharmaceuticals and related fields, although the majority of methods commonly used impose parametric assumptions on the distribution, which we wish to avoid (see Azzolina et al. (2021) for a full review of both parametric and nonparametric elicitation methods). We follow an existing belief elicitation approach in the literature that appears to work reasonably well, even with less educated populations. This involves providing respondents with a set of stones or beans, each representing probability units, and have them place them in a set of intervals, or bins, that cover the support of the distribution. For example, in eliciting priors over take-up, we gave respondents 20 stones, each representing 5 percent probability, and had them allocate them over a grid containing 20 intervals, each covering a range of 5 percentage points (see Figure D1 as an example).

Several practical considerations arise when using this method for outcomes with a potentially wide range of effect sizes. The first is that we do not want to have too many bins to overwhelm the choices of respondents, but at the same time we want to allow bins to be narrow enough in the likely range of parameter estimates that the prior is not degenerate. Further, in order to be able to easily aggregate priors across individuals, we wish the intervals to be the same for each respondent. We therefore used relatively wider intervals at the tails of the support of the possible distribution, and narrower intervals towards the middle (see Appendix Figure D2 as an example).

A further question that might arise when eliciting these distributions in the context of using them for a Bayesian impact evaluation is whether respondents will tell the truth, and if not, whether they should be incentivized with monetary payments to do so. In many field experiments, the results only occur far into the future, and so if priors are elicited at the time of launching the program, then payoffs may be so far into the future as to have little incentive effect on current responses. Instead, researchers may do better be ensuring the language used in eliciting priors ameliorates incentives to misstate priors. For example, following the approach commonly used in the corruption literature, we do not ask firms about their priors for the program's effects on themselves, but rather what they expect the effect of the program will be on average for all firms in the program. This both corresponds to the treatment effect we are estimating, as well as reducing the incentive for them to overstate what they think the program will do for them personally. These incentive effects may be more of a concern for some groups of respondents than others – for example, for policymakers rather than academics, which offers an additional reason to elicit priors from different groups. We see this as an interesting area for future research to examine the sensitivity of policymakers priors to their knowledge of how these

priors will be used in a Bayesian evaluation.[7]

3.3 Whose priors should be elicited?

We see benefits in eliciting priors from at least three potential sources of knowledge and beliefs about a program, and in obtaining priors from multiple people from within each source. The first group are the policymakers involved in designing and setting up an intervention. They are likely to have the best knowledge of the local context and of what the program is intended to do. Moreover, just as the process of pre-specifying outcomes in a pre-analysis plan can be useful for ensuring researchers and policymakers agree on what the program is intended to do, eliciting priors over these outcomes can help clarify where there is more or less heterogeneity in beliefs among different policymakers, and how certain policymakers are about different outcomes occurring. Posteriors based on their priors can then be most useful for policy decisions about the program. Second, eliciting the priors of academic experts is useful for incorporating the insights of existing expert opinion, and provides a way of combining their knowledge of the existing literature with how much they expect it to translate to the new context. Posteriors based on these priors can be useful for understanding how much academics should update their beliefs about the effectiveness of a type of program based on the data from this study. Third, program participants apply to programs with priors about how much the program will help them, and it is of interest to see how much they should update their beliefs about the program's effects after the study takes place.

We collected priors between June and October 2018, as the program was in the diagnostic phase. This timing was chosen such that the details of the intervention were as clear as possible, but so that it was before policymakers and firms would be able to see any program effects. Priors were collected from seven high-level Colombian policymakers involved in decision-making around the program, ranging from the vice Minister of Commerce to the program coordinator for the PEIP project. Academic priors were collected from eleven academics, all of whom had either published papers on management improvements or on Colombian firms. Firm priors were collected from the key decision-maker (typically general manager) at 10 of the firms in the treatment group.

There are two reasons for collecting priors from multiple respondents – in our case, from 7 to 11 respondents per group. The first is that, from the point of view of using the priors as a source of prior knowledge to inform the impact evaluation, the wisdom of crowds suggests that the average from the group may be more accurate and less noisy

[7]Dellavigna and Pope (2018) find no impact of offering financial incentives on the accuracy of MTurk forecasters. The incentives to deliberately misstate priors may be different if individuals think this will directly affect the allocation of resources, and then using mechanism design approaches as was done in Hussam et al. (2022) could be used.

than individual predictions. For example, Dellavigna and Pope (2018) find that taking the average of even five experts leads to a large improvement in accuracy over individual forecasts. Otis (2022) uses data from seven randomized experiments and finds that the average of groups of expert predictions does better than individuals at predicting which of two treatments will have larger effects, with only 10 experts needed to produce a 18 percentage point improvement compared to individual-level predictions. Aggregating across respondents then makes the results less sensitive to the idiosyncratic beliefs of any one individual, and can also provide a smoother distribution that is easier to fit a parametric model to (see below). Second, we are primarily interested in obtaining priors for types of users (policymakers, academics, program respondents), since there are typically multiple decision-makers involved in using knowledge.

3.4 Aggregation and fitting of elicited priors

Since respondents all used the same grid of bins to place their stones, we can easily aggregate responses by determining the proportion of all stones that get allocated to different intervals. This gives us an empirical CDF prior for each of the three groups of respondents. However, because we use Markov-Chain Monte-Carlo (MCMC) methods to simulate draws from unknown distributions for our Bayesian analysis, we require PDFs rather than CDFs in order to use these priors in our impact evaluation. We fit prior distributions to the elicited priors using two distinct types of parametric models: skew-normal distributions and finite mixtures of Gaussians with up to 5 components. We then judge the fit of these models by checking the fitted quantiles are close to their empirical counterparts. Full details of this procedure and our registered priors can be found as part of the study's registration in the AEA registry.

3.5 Literature-informed priors

We complement and compare our elicited priors to priors that are informed by the literature. Since there were no previous studies in the literature that conduct management improvement experiments aimed at improving exports, we can not use a meta-analysis of existing treatment effects as a prior. Instead, we use the result from McKenzie and Woodruff (2017) that an approximate estimate of the treatment effect of a business training intervention on firm outcomes is equal to the treatment effect on a business practices index, multiplied by the correlation between business practices and this firm outcome in the cross-section. Based on the existing literature on business training programs and management consulting, we took a literature-informed prior that the impact of the program on export practices would be a 0.10 increase, with a standard deviation of 0.03. We then multiplied this by the estimated coefficients in a baseline regression of our treatment

outcome variables on export practices, and assumed a Gaussian prior with mean equal to this product, and standard deviation derived from the standard error in the cross-sectional estimation and the assumed standard deviation on the impact on export practices. Full details and the registered priors are contained in the study's registration in the AEA registry.

3.6 Weakly Informative "Default" priors

For comparison purposes we also generate Bayesian results using highly diffuse or "weakly informative" priors, which are now the default standard for Bayesian analysts in cases where specific outside information is lacking (e.g. Meager (2019), Lemoine (2019), Thorlund et al. (2013), Chung et al. (2012)). These priors perform mild regularization on the estimation procedure, and we follow the literature above in using diffuse Normal priors centered around zero for coefficient parameters, and using diffuse half-Normal or half-Cauchy priors on variance parameters. In theory, the impact on the estimation from these priors should be quite minimal. This enables us to see how much of any difference in results is coming from the Bayesian impact evaluation techniques alone, versus through the specific additional information contained in our elicited priors.

4 Data and Methods for Estimating Treatment Effects

4.1 Data and Main Outcomes

Our primary hypothesis is that the program will lead more firms to export, diversity the range of products exported and destinations exported to, and improve the export performance of participating firms. Our primary outcomes of interest are therefore export outcomes. We use annual data on exports from 2010 to 2020 provided by the National Directorate of Taxes and Customs (DIAN) and supplied to us by the Colombian National Planning Department (DNP). 135 of the 200 firms exported at least once during these 11 years. These data provide export values at the 6-digit product and destination country level for each firm. For example, perfumes and cosmetics exported by a firm to Ecuador in 2018, or leather products exported by a firm to Chile in 2020. Using these data, we first measure the extensive margin of whether a firm is exporting at all, and then construct measures of export variety (number of countries, number of products, number of country-product combinations). We define export innovation as exporting a new country-product combination, and sum up the total value of exports. We merge these export data with data on formal employment provided by the Ministry of Health (the PILA), and use this

to construct an export productivity measure defined as exports per worker. Since exports and exports per worker are heavily skewed and contain many zeroes, our pre-analysis plan stated that we would take the inverse hyperbolic sine transformation of these outcomes. Finally, we also construct an overall export performance index, defined as the average of standardized z-scores of these different export measures. Appendix C defines each outcome in more detail.

We supplement these export and employment data with a combination of data from the program, a survey, and with linking the firms to other government datasets in Colombia. We use data from the application forms to describe the characteristics of firms at baseline, for stratifying the random assignment, and for balance checks. Program records provide information on take-up and usage of the intervention. A follow-up survey (described in section 5.2) is used to examine impacts of the program on export-specific and general management practices. Finally, in Appendix E we report impacts on secondary outcomes of interest such as sales, employment, survival, and productivity, by using data from 2018-2020 annual filings of firms in the Mercantile Registry (RUES) and to data from 2016 to 2019 in the Annual Manufacturing Survey (EAM). We did not elicit priors for these secondary outcomes, and so provide only frequentist and not Bayesian impact evaluation results for those outcomes.

4.2 Estimation of Treatment Effects using Frequentist Methods

Our frequentist estimation follows the approach standard in the literature. We use the following pre-specified Ancova linear regression specification to estimate the intention-to-treat effect. Our estimating equation for the ITT impact on outcome Y of firm i being assigned to treatment versus being assigned to control takes the form:

$$Y_{i,t} = \alpha + \beta Treat_i + \sum_{s=1}^{5} \gamma_s Y_{i,t-s} + \sum_{j=1}^{54} \delta_j 1(i \in strata_j) + \varepsilon_{i,t} \tag{4.1}$$

where $Y_{i,t-s}$ is the sth pre-intervention lag of the outcome of interest; δ_j are randomization strata fixed effects (following Bruhn and McKenzie (2009)); and β is the intent-to-treat effect. Robust (Eicker-White) standard errors are then used.

In addition to the standard hypothesis test that the average effect of being offered the treatment is zero, $\beta = 0$, we can also use the elicited priors to test additional hypotheses. We provide tests of the treatment effect being equal to the medians of the prior distributions of policymakers, academics, and firms.

The key assumption underlying equation 4.1 is the stable unit treatment value assumption (SUTVA). This will be violated if treated firms compete directly for export

sales with control firms, so that additional export success for the treated firms may come from competing away business from the controls. The sectoral heterogeneity of firms in our sample helps in this regard. Using the 6-digit product code, 62% of firms do not have a single other firm in the study exporting the same product as them, 70% do not have any other firm exporting the same country-product combination, and 94% have 3 or fewer firms exporting the same country-product combination. Moreover, while some of the more common 6-digit product codes are very specific (e.g. *uchuva* (cape gooseberry), *granadilla* (yellow passionfruit)), most of the more common product categories have more within-category heterogeneity (e.g. cotton t-shirts, long and short trousers for women and children, shirts and blouses of artificial or synthetic fiber, miscellaneous plastic products, miscellaneous steel products). Our assumption is therefore that any export growth from the treated firms is unlikely to be primarily business stealing from control firms.

4.3 Estimation of Treatment Effects using Bayesian Methods

Our headline Bayesian analysis takes as its starting point the regression in equation 4.1 as the conditional mean of the outcomes of interest, constructs a likelihood around this regression, and then adds priors. This modelling procedure has several distinct components worth explaining in detail.

The first task is to place a likelihood on the regression errors $\varepsilon_{i,t}$, which corresponds to placing a likelihood on $Y_{i,t}$ given the covariates. In our main specification, we use a Gaussian likelihood for all outcomes as this corresponds most closely to the Ordinary Least Squares estimation approach on the linear regression model, because the point estimates for the regression coefficients from MLE on the Gaussian likelihood are identical to the OLS point estimates.[8] This choice means that when prior information is weak our Bayesian models ought to default to the "standard" frequentist inference on the coefficients delivered by fitting OLS to the regression model above (with a caveat that even weak prior information could theoretically be useful and influence the inference in certain cases). We specify a single variance parameter σ^2 for parsimony; this should not make much difference to the inference on the treatment variable in an RCT, since $Treat_i$ and $\epsilon_{i,t}$ are fully independent and not just mean-independent. This produces the following likelihood model:

$$Y_{i,t} \sim \mathcal{N}(\alpha + \beta Treat_i + \sum_{s=1}^{5} \gamma_s Y_{i,t-s} + \sum_{j=1}^{54} \delta_j 1(i \in strata_j), \sigma^2) \tag{4.2}$$

We now need to place priors on each of the parameters that govern the likelihood. We

[8] This is because the OLS objective function is the kernel of the Gaussian Likelihood with respect to β.

perform the analysis using several different priors on the key parameter of interest β, as follows:

1. Academics' Priors
2. Policymakers' Priors
3. Firms' Priors
4. Literature Priors
5. Weakly Informative "Default" Priors

The first three categories of priors are our elicited priors, discussed extensively in section 3; the literature priors are discussed in section 3.5 and the default priors are discussed in section 3.6. As a quick refresher, we note that the elicited priors take the shape Skew-Normal or a mixture of up to 5 Normals depending on what fits the expert prior data best, while the literature priors and default priors are Normal.

The next set of parameters to consider are the control variables, and here we confront a broader issue for translating the frequentist approach above into a Bayesian model, because we have many controls due to the 54 δ_j strata fixed effects. Our motivation to perform stratified randomization was to ensure that balance holds along key firm characteristics in the finite sample; ex-ante, across repeated experiments, this improves precision in the estimation. In a Bayesian perspective, it is standard to consider the ex-post precision, conditional on the data and the model, since that is the variance one actually has in any given analysis. Regression models with large numbers of strata fixed effects can be problematic even for OLS when the overall sample size is small, especially when the strata we draw from are very small (Athey and Imbens (2017)). Moreover, unlike OLS, Bayesian estimation strategies do not in general "partial out" uncertainty on additive parameters because Bayesian inference is done jointly across all parameters; high posterior variation on δ_j can therefore propagate to higher uncertainty on treatment effects. Another way to understand the need for an adjustment to the estimation is to note that with 200 data points and 54 strata fixed effects, overfitting the sample data is a real possibility.

The standard Bayesian modelling approach to handling large numbers of nuisance parameters without overfitting is apply some form of regularization (Gelman et al. (1995)). We place a hierarchical model on the strata coefficients in order to shrink them to their shared mean, and then we regularize this mean towards zero. Specifically, we add to the likelihood the structure that $\delta_j \sim N(\delta, \sigma_\delta^2)$. We then place priors on these new "hyperparameters" $(\delta, \sigma_\delta^2)$ which are Normal and half-Normal respectively, both centered

at zero. We use prior standard deviations equal to 25% of a crude estimate of the outcome data's scale for the hypermean δ and 50% for the standard deviation σ_δ, in order to allow for more heterogeneity if the data suggests it. This structure regularizes the estimates in a similar manner to a Ridge regression penalty (see Hastie et al. (2009) for the exact relationship). As with Ridge, if the data strongly suggests any of these interactions are important in predicting the outcomes in ways beyond their correlation with the treatment assignment variable, they will be able to overcome the penalty imposed by the hierarchical structure and the priors. Note however that we do not constrain the role of the 3 main factors that drive the stratification: firm size (implemented via 3 categories), exports in last 3 years versus not, the export practices index and an indicator for taking extreme values of any of the baseline outcomes, as these are potentially important controls in their own right and are more likely to improve model fit overall.

Finally, we need priors on the other control variable coefficients and other parameters in the likelihood. We generally use reasonably weak, diffuse priors for these but we do incorporate genuine prior information where possible. For α we use a weakly informative Normal prior centered at the baseline average of the outcome and having the same scale as the baseline variance. For the lagged outcome coefficients γ_s, as well as the 3 central covariates that drive our stratification design, we use a diffuse Normal centered at zero with a very large scale of variation. For σ^2 we use a very diffuse half-Normal bounded below at zero. These choices correspond to the "weakly informative" type of default prior approach seen in previous work, as discussed in section 3.6.

Once we have the likelihood and priors, to proceed with Bayesian inference, one multiplies the likelihood by the priors which yields the posterior distribution which captures the joint uncertainty present about all unknown parameters conditional on the sample of data we have. In practice, this full joint posterior distribution function is typically challenging or even impossible to compute analytically as it is not of a known, standard functional form. The shape of this joint posterior distribution function can be approximated using Markov Chain Monte Carlo (MCMC) Methods. We use a Hamiltonian Monte Carlo algorithm in which tuning parameters are automatically disciplined using the No-U-Turn-Sampler, implemented via the software package Stan in R. In each case we examine the performance of the HMC via the effective sample size, traceplots and R-hat criterion. For more information about these metrics and related computational issues see the Appendix of Meager (2019).

The fundamental output of Bayesian analysis is this joint posterior over all parameters, but our primary interest is inference on the impact of the program. Thus, we integrate out the marginal posterior distribution on β from each analysis and report the central 95% credible interval as our headline Bayesian result. We can also compute other quantities of interest, such as the probability that β is large enough for the program to pass a cost-

benefit analysis; we also compute and report these results in section 5.4. In addition, once we have specified a proper likelihood on the data, it is natural to consider jointly modelling the potential outcomes to get more detailed causal inference on the distribution of treatment effects in the style of Imbens and Rubin (2015) – we pursue this as an extension to our main analysis (additional details and results provided in section 5.4).

We do note that the approach described above differs from the way that Bayesian statisticians or analysts in other fields would tend to approach this same data. Modern Bayesian data analysis differs from the Frequentist econometric approach in more respects than just the addition of priors; in particular, Bayesian analysis typically uses richer likelihood models that are tailored to the specific functional forms of the outcome variables in question, since much is typically known about them (e.g. productivity is bounded below at zero and thus likely right-skewed, export probability is binary, etc). However, we pursue a Bayesian analysis using Gaussian likelihoods because in the absence of priors these likelihoods deliver the OLS estimates of the regression coefficients, making it easier to compare the frequentist and bayesian results and to understand the role of our informative expert and firm priors.[9]

5 Results

5.1 Take-up and usage

Take-up and usage of the program was high, especially given the requirement for firms to pay US$1,000-$2,000 to participate. 88 of the 100 firms assigned to treatment are recorded in the administrative data as starting the intervention, and 83 have positive hours of consulting recorded. Figure A1 shows that this take-up rate is high compared to the medians of the prior distributions for take-up of the participating firms (58%), academics (63%) and policymakers (73%), and in line with the mean of our pre-specified literature prior (80%). Moreover, of the 83 firms with recorded hours, 94% (all but 4 firms) received at least 100 hours of consulting, with the median firm that took up the program receiving 195 hours. Appendix Figure B1 shows the distribution of total hours of consulting, and we see that half of all firms actually received slightly more than the 190 hours initially promised by the program.

Table 2 disaggregates the consulting received by area. Of the 83 firms with recorded activity data, 72 had hours in all three areas of consulting, 8 in two areas, and 3 in only one area. The most common two areas were the compulsory commercial strategy area,

[9]There are also substantial challenges with implementing the analysis with the richer likelihoods, which we document in Appendix G.

which 79 firms[10] received for a median of 35 hours, and operational productivity, which 71 firms chose and received for a median of 80 hours. Most firms in the program therefore received these two areas, and then were divided in their choice of a third area, with 34 receiving labor productivity, 32 quality, and 19 energy consulting.

5.2 Impacts on business and export practices

The most immediate impacts of management consulting interventions are typically seen on management practices, as firms implement the changes recommended by consultants. We hired Innovations for Poverty Action to conduct a follow-up survey of firms between November 2019 and May 2020. This was able to collect data on 172 of the 200 firms (89 treatment and 83 control), and additionally confirm that 3 other firms were closed (2 control, 1 treatment). We measure impacts on 15 different export-oriented management practices, such as whether the firm participates in trade fairs, has received a quality certification for an export market, does direct marketing to international customers, or had received information about distributors abroad. In addition, we measure impacts on each of 40 different general management practices that cover the five areas of consulting, divided into operations practices (10), labor productivity practices (9), quality practices (6), energy practices (4), and commercial practices (11). Appendix C provides definitions of all measures, which were developed with feedback from the different consulting entities to ensure they covered the main areas of advice given. Our main measures of export practices and management practices are then defined as the proportion of these practices that are used by firms.

Figure 2 shows the estimated intention-to-treat effects of being assigned to consulting on these two indices of management and export practices, as well as on the five sub-components of the overall management practices index. We did not elicit priors on these outcomes, so only show frequentist point estimates and confidence intervals from estimating equation 4.1.[11] The program had no significant impact on our measure of export-specific management practices. On average, control group firms were using 59 percent of the export practices, and the estimated change of -4 percentage points is small and not statistically significant. Appendix Figure C1 also shows no significant impact on any of the 15 individual practices that make up this overall index.

In contrast, there was a significant improvement in general management practices. The control group were using 64 percent of these practices, and there is a significant

[10] Although this was a compulsory area, because it started several months after operational productivity and labor productivity, 4 firms started activities but dropped out before they began their commercial strategy.

[11] We control for a baseline measure of export practices or general management practices in these regressions

treatment effect of 6.7 percentage points. Examining the sub-components of this overall index, we see the biggest changes were in the two areas in which more of the firms received consulting. Operations practices improved 11.9 percentage points from a control mean of 71 percent, and commercial practices improved 7.9 percentage points, from a control mean of 66 percent. The changes in labor productivity, quality, and energy practices are small in magnitude and not statistically significant. Appendix Figure C2 examines in more detail the individual practices in the operations and commercial practices indices. We see that the largest improvements in operations practices occur in the use of lean manufacturing methods: using VSM, 5S, and continuous improvement methods, with a significant improvement also in communicating strategic goals around operations. The largest improvements in commercial practices occur in practices to better understand and connect with customers, through setting up a CRM system and doing market research on customers.

As noted in section 2.3, the theory of change offered two potential pathways through which the program might improve exports: a direct pathway through improving export-specific practices, and an indirect pathway through improving general management practices. The program does not appear to have had any impact through this direct channel, but has improved processes through this indirect channel. We next look to see whether this was sufficient to increase exports.

5.3 Impacts on export outcomes

Figure 3 plots the trajectory of means for the different export outcomes by treatment status over the period 2010 to 2020. We see that firms were on an upward export trajectory prior to participating in the program, reflecting that the program selected firms that were exporting or planning on starting to export. The treatment and control groups track each other closely prior to the program, as would be expected with random assignment. We then measure treatment effects in two post-treatment years: 2019 and 2020. Firms were still receiving the intervention for the first half of 2019, while in 2020 the world experienced the COVID-19 pandemic that placed restrictions on travelling, led to some temporary closures, and impacted both the demand for exports and the ability to export. Visually, it appears that the treatment and control groups again look similar to one another in 2019, and diverge somewhat in 2020, with exports falling in the treatment group and rising in the control group.

Table 3 then provides our frequentist estimates of the ITT effects on these different export outcomes, after controlling for randomization strata and pre-treatment lags as per equation 4.1. For each outcome we report the estimated treatment effect in 2019 and in 2020, and then as well as testing that the treatment effect is zero, also test the

null hypotheses that the treatment effects are equal to the medians of the different prior distributions elicited.

Consider first the impact on the extensive margin of whether firms are exporting at all, shown in panel A of Table 3. 54 percent of the control group exported in 2019 and 57 percent in 2020. The estimated treatment effect in 2019 is a statistically insignificant -0.5 percentage points (p.p.), with a 95 percent confidence interval of [-9 p.p., +8 p.p.]. In 2020, the estimated treatment effect is a statistically insignificant -6.5 p.p., with a 95 percent confidence interval of [-15 p.p., +2 p.p.]. Although we cannot reject the null hypothesis that the treatment effect is zero, we can reject that the treatment caused exports to increase as much as expected by the median of the literature, firm, and policy priors (9 to 13 p.p.), and also in 2020 that the effect as large as the median of the academic prior (6 p.p.).

Figure 4 then shows the results of our Bayesian impact evaluation on this outcome ("export at all"). The frequentist confidence interval is shown at the bottom of the figure in red, and we show intervals that cover the 2.5th to 97.5th percentiles of the different prior distributions in light blue, and then 95 percent coverage intervals from the estimated Bayesian posterior distributions using each prior. First, note that the posterior coverage intervals using a non-informative (diffuse) prior are extremely similar to the frequentist confidence intervals, showing that the Bayesian estimation methods per se are not changing our results. We then examine how these intervals change when we bring in informative priors. We see that the posterior coverage intervals for the three types of elicited priors (policy, firm, and academics) almost completely overlap with the frequentist intervals. That is, the signal in the data is strong enough relative to these priors that we almost completely update these priors towards the data. A graphical display of how the priors are updated for academics is provided in Appendix F, figure A8. In contrast, the literature prior for this outcome was for a larger and more precise positive effect, and the posterior based on this prior moves a lot towards the data, but does not fully update.

Next, panels B, C, D and E of Table 3 examine impacts of the program on the number of products exported (control mean 4.2 in 2019, 4.1 in 2020), the number of countries the firm exports to (control mean: 2.3 in 2019, 2.4 in 2020), the number of unique product-country combinations exported (control mean: 10.1 in 2019, 10.0 in 2020), and whether the firm has exported a new product-country combination (export innovation) (control mean 0.40 in 2019, 0.41 in 2020). The treatment impacts are close to zero, and not statistically significant for all of these four outcomes in 2019. The 2020 impacts are more negative than those in 2019, and in the case of the number of product-countries, the negative effect is statistically significant. In all cases the estimated treatment effects are smaller than the medians of the different prior distributions, and, in 2020, we can reject all the alternative nulls that the treatment effects equal these medians.

Figure 5 provides the associated Bayesian impact evaluation results for export variety and Figure 6 impacts on export innovation.[12]. We see that the posterior coverage intervals again show substantial updating of the priors towards the data. An interesting illustration of the value of informative priors comes from looking at the estimated impact of the program on the number of product-country varieties exported in 2019. The 95 percent frequentist confidence interval is [-1.9, +1.9] products, giving an interval width of 3.8. The elicited firm priors were for a small, positive effect, with a 95 percent interval of [-1, 9.6]. Bringing in this informative prior which is consistent with the data narrows the posterior interval to [-0.8, 1.8] products, for an interval width of 2.6. The policymakers had even narrower priors, with a prior interval of [0.5, 5], and using this prior results in a posterior interval of [-0.03, 2.1] products, giving an interval width of 2.1. That is, when informative priors are largely consistent with the data, using Bayesian impact evaluation offers the possibility of more precise estimates of the treatment effect than we would get using the data alone. However, in 2020, where the data shows a more negative impact, these stronger priors mean that the posterior intervals only partially update towards the data.

Export value and export labor productivity are highly skewed outcomes with a mass of observations at zero and an extremely long tail. For example, in 2019, 94 of the 200 firms had zero exports, median exports was $5,029, mean exports $396,000, the standard deviation is $1.2 million, the 90th percentile $1.1 million, and the 99th percentile $6.7 million. Moreover, theoretically it seems more likely that treatment would lead to a similar percentage increase in exports for all treated firms than a similar level increase. Our pre-analysis plan therefore specified that we would follow a standard approach in the literature of using the inverse hyperbolic sine transformation of these outcomes, but noted that power was still likely to be low, with an anticipated minimal detectable effect of a 49 percent increase in export value. The hope was that bringing in prior information may enable us to obtain narrower posterior intervals if these priors were consistent with the data. We also placed 19 of the firms with highest baseline export performance in a separate strata, in order to examine how sensitive the results are to these tail firms. Table E1 shows dropping this strata lowers the control means, but leads to only small changes in the estimated treatment effects.

Panels F and G of Table 3 show that the estimated treatment effects on export value and export productivity are negative and large in magnitude, but with considerable uncertainty attached. The impacts are not statistically significant in 2019, but are statistically significant in 2020. Figure 7 provides the Bayesian impact evaluation results. With a diffuse prior, the Bayesian credible posterior intervals again are similar to the frequentist confidence interval. However, with informative priors, we see very little updating of the

[12]Figure D1 shows the results for the number of products and number of countries

elicited priors take place. The data here are not very informative about the impact of the treatment on these outcomes, and so the posteriors place much more weight on the priors than the data. Two exceptions are the case of the literature priors, where our prior based on the literature has a wide distribution, reflecting the difficulty in estimating impacts on these outcomes; and the academic priors for impacts on export labor productivity, which are also quite wide. In these cases, a bit more updating takes place, since the priors are more diffuse.

Our overall index of export performance takes the average of standardized z-scores of all these different export outcomes, and provides a summary measure of the impact of the program on exporting. Panel H of Table 3 shows the overall impact is a statistically insignificant -0.012 standard deviations in 2019, and a statistically significant -0.112 standard deviation effect in 2020. That is, taken together, the results show the program actually reduced exporting. We thought that it might be hard for firms to understand impacts on a summary index in terms of standard deviations, so did not elicit priors from firms for this aggregate outcome. Figure 8 shows the Bayesian results using the literature, policymaker and academic priors. We see substantial updating of the priors towards the data, with the posterior intervals centering around zero in 2019 and around small negative impacts in 2020.

These findings of a reduction in exporting are consistent with the responses from directly asking firms about their export strategies in our follow-up survey. We asked firms whether they had attempted to export to a new destination, even if not yet successful, and the control group (56.6%) was more likely to have attempted exporting than the treatment group (48.3%). We also asked how the company's focus on exports had changed in the pat 12 months, and 23.6% of the treatment group said they had become less focused on exports, compared to 18.1% of the control group. Treated firms are also more likely to now say their main growth focus is on the domestic market (50.6%) compared to control firms (43.4%). We analyse these and several other secondary outcomes and examine robustness to dropping particularly outlying firm strata in appendix E.

5.4 Bayesian Decision Analysis

The Bayesian approach offers several additional possibilities for analysis to directly inform policy decisions. In this section we will consider and execute two of these options. First, we compute the probability that the average effects of the intervention pass a minimal cost-benefit threshold, given the evidence. Second, as decisions may be based on more than just average effects, we derive and estimate the conditional distribution of individual treatment effects using additional priors on the correlation between firms' potential outcomes, an analytic approach set out by Imbens and Rubin (2015).

First, consider assessing the probability that the program's average effects pass the minimum threshold necessary for policymakers to consider scaling it up. While this is sometimes attempted in frequentist analysis using the sampling distribution of the estimator, there is no epistemic justification for doing so, as this distribution is an asymptotic property of the estimator in repeated samplings, not the uncertainty we have about the estimand itself. In the Bayesian framework, the posterior distribution directly captures the probability that parameters take certain values conditional on the data and priors - this distribution can be used to make probability statements about the treatment effect. During our prior elicitation exercise with the policymakers we interviewed, we also asked them the minimum effect size that would make it worth scaling up the program as a broader policy. The marginal posterior distributions on the treatment effect β for each outcome from our analysis in section 4.3 capture the probability that the effect takes any given value, which allows us to directly compute the probability that the effect passes this threshold. Because we have used Markov-Chain Monte Carlo methods to characterize the posterior distributions, and these methods simulate draws from these posteriors, to then compute the desired probability one simply computes the proportion of draws from the posterior that lie above the threshold.

The results of this exercise for the policymaker, academic and firm priors, plus literature priors and default priors, are shown in table 4 for all outcomes for both 2019 and 2020. Overall, the probability of attaining the minimum viable threshold for scaling up the policy is very low across all outcomes and priors; typically much less than 10% chance, and in some cases, virtually zero chance. The chances are typically lower in 2020 than in 2019, as one would expect given the movement of the data towards more negative outcomes. The exception to this pattern is export value and productivity, where chances can be much higher - but this is primarily because the data contains little information about these outcomes and as a result we have little updating away from the priors even in 2020. Overall we find the data much more informative on the other five outcomes, all of which show very little chance that the effect is large enough to merit scaling up the policy.

Second, as we recognise that decision-makers may wish to go beyond the average effects, we use Bayesian modelling of potential outcomes to infer the distribution of treatment effects as laid out in Imbens and Rubin (2015). Given a likelihood model for the distribution of potential outcomes, and prior distributions the unknown parameters, one can derive the conditional posterior distribution of the treatment effect estimand given the randomized assignment and subsequent observed data. This distribution can then be used to obtain the posterior distribution of individual treatment effects in the sample. This exercise can not be done without a prior on the correlation in the potential outcomes because the data never contains information about this correlation, since by definition we

never observe both $Y(0)_i$ and $Y(1)_i$ for any i.

Model-based inference on potential outcomes with covariates X_i requires a joint liklihood $f(Y(0)_i, Y(1)_i|X_i)$. Our previous analysis – linear regression of outcomes Y_i on X_i via ordinary least squares – corresponds to marginal distributions of $Y(0)_i$ and $Y(1)_i$ which are both Gaussian distributions with mean $X_i\beta$ (using generic econometric notation) which differs only in switching on or off the binary treatment indicator covariate. This was discussed already in setting up the likelihood in equation 4.2. To turn this set-up into a joint likelihood model requires structure on the covariation between $(Y(0)_i, Y(1)_i)$. One parsimonious and tractable such structure, as noted in section 8.4 of Imbens and Rubin (2015), is the bivariate Gaussian. Suppose then that $f(Y(0)_i, Y(1)_i|X_i)$ is a bivariate Gaussian with respective marginal variances $(\sigma(0)^2, \sigma(1)^2)$ and correlation parameter ρ. For brevity we will now denote the conditional expectation of Y_i given only the non-treatment covariates by $X_i\pi = \alpha + \sum_{s=1}^{5} \gamma_s Y_{i,t-s} + \sum_{j=1}^{54} \delta_j 1(i \in strata_j)$, and we will assume that the treatment does not affect the variance of the outcome, so $\sigma(0)^2 = \sigma(1)^2 = \sigma^2$. Then, by the properties of multivariate Gaussians, the distribution of the *missing* potential outcome for any firm i conditional on seeing the realised potential outcome is:

$$Y(1 - Treat_i)_{i,t} \sim \mathcal{N}(\beta(1 - Treat_i) + X_i\pi + \rho(Y_i - \beta Treat_i - X_i\pi), (1 - \rho^2)\sigma^2) \quad (5.1)$$

This is analogous to equation 8.34 of Imbens and Rubin (2015). Drawing the missing potential outcomes from this distribution, conditioning on the posterior draws of the parameters from our Bayesian model, allows us to compute the individual treatment effect $\tau_i = Y(1)_i - Y(0)_i$ for each firm. The posterior uncertainty on each τ_i comes from the need to infer the parameters that govern the likelihood, and then impute the missing potential outcome. It is instructive to examine the distribution of τ_i across all firms in the sample to understand the potential for heterogeneous effects. These may arise even though β is just a number, because of correlations in the unmodelled variation in the potential outcomes; that is, when ρ is not zero.

Any inference - frequentist or Bayesian - on the distribution of individual treatment effects (ITEs) in the sample requires some information about ρ. As the data cannot speak to the value of this parameter, it is natural in the Bayesian set-up to use priors. We elicited beliefs about ρ from academics and policymakers using the same methods as used when we elicited priors about β. Both priors on rho were largely positive; that is, both academics and policymakers believed that firms doing well under the control regime (the status quo) would also do well under treatment. Combining our elicited priors on ρ and other parameters with the model in equation 5.1 allows us to infer the distribution of individual treatment effects (ITEs) with appropriate posterior uncertainty on each ITE.

The results for the Index outcome data are shown in figure 9 for both 2019 and 2020 for the policymaker priors These figures show that the majority of the firms have ITEs centered at zero, with limited potential for heterogeneous effects in both 2019 and 2020. There is a reasonably high probability that at least some firms experienced positive effects, but the same is true of negative effects, and most firms have ITEs estimated to be almost exactly zero. Both the spread of effects and the uncertainty in these tails is more pronounced in 2020 relative to 2019. Overall however it seems the majority of the firms in the sample were unlikely to have seen any results from the intervention, and while some are likely to have benefited, others are likely to have experienced negative effects. This is most likely a result of combining the data with a prior positive correlation on the potential outcomes (so firms who did well in treatment would also have done well in control, and vice versa). We obtain similar results with the academics' priors, shown in the additional results Appendix F, figure A9.

Given the somewhat negative effects on average in 2020, it may be surprising that there is still some potential for positive tail effects for other firms. This result may be an artefact of the likelihood model chosen: the Gaussian enforces symmetric tails in the outcomes, which propagates to a Gaussian and thus symmetric distribution of individual treatment effects. The Gaussian may be a particularly constraining assumption in our case especially because the outcome data is not symmetric – it has a large discrete spike at zero and then a positive continuous mass situated away from zero. While we could build a likelihood model to accommodate this, in our case the priors on treatment effects will have a one-to-many mapping to priors on parameters of such a model (because a treatment effect could be achieved by moving firms out of the spike, or from changing the distribution of outcomes for firms already out of the spike in the control state). Indeed, this is a general issue for our Bayesian analysis preventing the use of more complex likelihoods without substantial additional effort; further details of this problem are discussed in appendix G. Overall, however, we still find the analysis informative, and it is worth noting that the results of the current approach do suggest only a limited potential for heterogeneous effects.

5.5 Why were impacts not as large as expected?

We see that, contrary to the priors of academics, policymakers, the participating firms, and the existing literature, the program did not succeed in increasing exports, and in fact appears to have actually reduced exporting in 2020. We believe that these results stem from a combination of the type of consulting advice given, the intensity of the intervention, and the quality of this advice. In addition, the heterogeneity of the firms in the program makes it harder to detect impacts, and likely made it harder to offer a standardized program.

A first issue was the type of consulting advice provided. As we discussed in section 2.3, there are two ways we viewed the consulting as potentially changing exports: a direct approach in which it improved export-specific management practices and really focused on customers and markets abroad, and an indirect approach in which it aimed to improve general management practices and productivity, enabling firms to be more competitive and (perhaps eventually) better able to compete in foreign markets. The government's focus over time shifted to this second approach, and we see that the program was not successful in changing export-specific management practices, but only general management practices. Our qualitative discussions with the commercial strategy consultants and several program firms found that the consultants believed that the majority of firms could not consistently produce at the capacity and quality needed to go to external markets, and so, in some cases, actually recommended that they focused more on the domestic market than foreign markets. Appendix B does give some examples of specific advice offered on particular export matters, but for the most part, the commercial advice was not specifically tailored to exporting, and the focus in the other four consulting areas was more on general productivity than what was needed for overseas export markets.

This might suggest that we should expect to see improvements in other firm outcomes, even if it does not immediately translate into export performance. Appendix E uses data on firm survival, employment, sales, profits and productivity to examine whether this improvement in general management practices improved other firm outcomes. We see a marginally significant 5 percentage point improvement in firm survival rates by the end of 2020, consistent with these general management improvements potentially helping firms survive through the COVID pandemic. However, these additional firms that survived were smaller and less productive ones, and so this does not help export outcomes. Moreover, although the heterogeneity in firms makes it hard to measure impacts on these firm performance outcomes, there is certainly no strong evidence of improvement in these other domains. Employment impacts are small (a statistically insignificant 0.4 worker change relative to a sample mean of 80 workers), and the point estimates for the treatment impacts on sales, profits, and sales per worker are all negative, with the -12.7 percent reduction in sales in 2020 statistically significant. So it does not appear that firms are well on their way to increasing exports through this indirect route of increasing domestic sales and productivity first.

This raises the issue of the quality and intensity of the consulting provided. We surveyed the consultants to understand their backgrounds. The median consultant was 45, with 17 years of experience, and a Masters' degree, and 73 percent said they had previously worked for a multinational or an exporter. So, on paper at least, the consultants had relevant expertise. However, the consulting intervention was shorter in intensity and less coordinated than in some other consulting experiments. Firms received 190 hours of

consulting, including only 30 on commercial strategy, and this was divided among three consulting companies who operated largely separately from one another, and who each spent a lot of time in diagnosis. Since consulting firms had been hired to work solely on one area, their main focus was on measuring indicators and helping firms in that one area, without regard to whether this change would be the most likely to deliver improvements in overall productivity or in the ability of the firm to export. In contrast, the Technological Extension project in Colombia studied by Iacovone et al. (2022) gave firms 500 hours of individualized or 408 hours of group-based consulting. In those cases the consulting across multiple areas was delivered in a coordinated fashion by a single Colombian consulting organization. This resulted in a 4-14 percent improvement in productivity. In the Indian study of Bloom et al. (2013), textile firms received 781 hours of consulting from an international consulting firm and obtained a 17 percent improvement in productivity.

Moreover, in contrast to these other studies, the firms in this experiment were much more heterogenous in size and industry. As well as making it harder to detect impacts, this heterogeneity likely also affected the quality of the intervention. Our qualitative interviews with firms revealed complaints that the consultants did not have knowledge on the specific industries firms were operating in. For example, a fashion firm noted they spent part of their time educating the consultants on their industry. This is likely to have been particularly a problem for efforts to advise firms on commercial strategy and new export markets, since the consultants would not know specifics of industry trends, quality standards, nor of relevant overseas buyers. Instead, consultants appear to have focused almost solely on narrower applications of lean manufacturing practices, which might have reduced bottlenecks, but which were not clearly linked to sales strategies. We explore robustness of our main results to dropping particularly outlying firm strata (in terms of high exporting) in appendix E.

It is possible that some of these changes will take longer to manifest themselves in firm productivity, and that the COVID-19 pandemic slowed this process down further. It is therefore theoretically possible that the program may have stronger impacts over a longer term period, as was the case of the consulting intervention in Bruhn et al. (2018). But at least in the first two years after the intervention, there is no sign of the program improving exporting through either the direct or indirect channels.

6 Conclusions

Policymakers, academic researchers, and Colombian firms all had priors that consulting services would lead to an increase in firm export outcomes. We have shown how these priors can be incorporated into a Bayesian impact evaluation. Contrary to these priors, our experiment finds no significant impacts on exporting in 2019, and some evidence of

a fall in exports in 2020. Our range of different export outcomes illustrate the different roles informative priors can play in analysis. Some of our outcomes, like the extensive margin of whether firms export at all, are estimated relatively precisely, and our Bayesian posterior credible intervals overlap with frequentist confidence intervals in these cases, fully updating the priors given the data. Here, having the priors helps communicate to a policy audience that, despite the results disagreeing with their prior beliefs, they should strongly update those beliefs given the data. For a second set of outcomes, such as whether firms introduce new export varieties, the priors were for a very small effect, which is consistent with the data, and using these priors results in Bayesian credible intervals that are narrower than a standard frequentist confidence interval. This shows the potential for Bayesian impact evaluation to lead to more precise results when priors and the data align. Finally, for our export value and export labor productivity outcomes, the data are very noisy, and the priors are not updated very much given these data. This illustrates that different interest groups may not want to update their priors about an intervention very much if the experimental results are not very precise.

Our analysis highlights some of the practical issues involved in using informative priors in practice. These issues include the need to deal with incomplete compliance and be clear on which treatment effect is being estimated; questions about who to elicit priors from and how to gather and fit these priors; and issues that arise when the number of randomization strata or controls are relatively large compared to the sample size. Our results appear to be reasonably robust to many of these choices, and in particular, are similar regardless of which set of informative priors we use. This should alleviate concerns that the priors of any one individual drive the results. By comparing our estimates to those using a non-informative (default) prior, we also see that it is the inclusion of informative priors, and not other Bayesian modeling choices like regularization of strata fixed effects, that drives results.

We see the promise of the Bayesian approach with informative priors as particularly useful for expensive long-term experiments that have a limited sample size or potential concerns about statistical power. Many experiments with small and medium enterprises, schools, and health clinics take this form. The Bayesian approach could also prove helpful in the estimation of treatment heterogeneity: Gelman (2018) shows the sample sizes needed to detect treatment interactions can be 16 times that of those needed for main effects, making most studies underpowered for heterogeneity. Incorporating informative priors about the extent and dimensions of heterogeneity could help overcome these power problems. As such, we see high potential for informative priors to be used as part of Bayesian impact evaluation in many future experiments.

References

Abadie, A. (2020): "Statistical Nonsignificance in Empirical Economics," *American Economic Review: Insights*, 2, 193–208.

Andrews, I. and J. Shapiro (2021): "A Model of Scientific Communication," *Econometrica*, 89, 2117–2142.

Athey, S. and G. W. Imbens (2017): "The econometrics of randomized experiments," in *Handbook of economic field experiments*, Elsevier, vol. 1, 73–140.

Atkin, D., A. Khandewal, and A. Osman (2017): "Exporting and Firm Performance: Evidence from a Randomized Experiment," *Quarterly Journal of Economics*, 132, 551–615.

Azzolina, D., P. Berchialla, D. Gregori, and I. Baldi (2021): "Prior Elicitation for Use in Clinical Trial Design and Analysis: A Literature Review," *International Journal of Environmental Research and Public Health*, 18.

Bloom, N., B. Eifert, A. Mahajan, D. McKenzie, and J. Roberts (2013): "Does Management Matter? Evidence from India," *Quarterly Journal of Economics*, 128, 1–51.

Bloom, N., K. Manova, J. V. Reenen, S. Sun, and Z. Yu (2021): "Trade and Management," *Review of Economics and Statistics*, 103, 443–460.

Bloom, N., R. Sadun, and J. V. Reenen (2016): "Management as a Technology," *Stanford Working Paper*.

Breinlich, H., D. Donaldson, P. Nolen, and G. Wright (2017): "Information, Perceptions and Exporting – Evidence from a Randomized Controlled Trial," *University of Nottingham Working Paper*.

Bruhn, M., D. Karlan, and A. Schoar (2018): "The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico," *Journal of Political Economy*, 126, 635–687.

Bruhn, M. and D. McKenzie (2009): "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 1, 200–232.

Cadot, O., A. Fernandes, J. Gourdon, and A. Mattoo (2015): "Are the benefits of export support durable? Evidence from Tunisia," *Journal of International Economics*, 97, 310–324.

Casey, K., R. Glennerster, and E. Miguel (2012): "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan," *Quarterly Journal of Economics*, 127, 1755–1812.

Chung, Y., S. Rabe-Hesketh, A. Gelman, J. Liu, and V. Dorie (2012): "Avoiding boundary estimates in linear mixed models through weakly informative priors," *U.C. Berkeley Division of Biostatistics Working Paper Series*.

Coffman, L. and M. Niederle (2015): "Pre-analysis Plans Have Limited Upside, Especially Where Replications Are Feasible," *Journal of Economic Perspectives*, 29, 81–98.

Custódio, C., D. Mendes, and D. Metzger (2020): "The impact of financial education of executives on financial practices of medium and large enterprises," *Imperial College London Working Paper*.

Delavande, A., X. Gine, and D. McKenzie (2010): "Measuring Subjective Expectations in Developing Countries: A Critical Review and New Evidence," *Journal of Development Economics*, 94, 151–163.

Dellavigna, S., N. Otis, and E. Vivalt (2020): "Forecasting the Results of Experiments: Piloting an Elicitation Strategy," *AEA Papers and Proceedings*, 110, 75–79.

Dellavigna, S. and D. Pope (2018): "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy*, 126, 2410–2456.

Dellavigna, S., D. Pope, and E. Vivalt (2019): "Predict science to improve science," *Science*, 366, 428–429.

Gelman, A. (2018): "You need 16 times the sample size to estimate an interaction than to estimate a

main effect," *Stat Modeling Blog*.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995): *Bayesian data analysis*, Chapman and Hall/CRC.

Girma, S., Y. Gong, H. Görg, and Z. Yu (2009): "Can Production Subsidies Explain China's Export Performance? Evidence from Firm-level Data," *Scandinavian Journal of Economics*, 111, 863–891.

Groh, M., N. Krishnan, D. McKenzie, and T. Vishwanath (2016): "The Impact of Soft Skill Training on Female Youth Employment: Evidence from a Randomized Experiment in Jordan," *IZA Journal of Labor and Development*, 5.

Hastie, T., R. Tibshirani, J. H. Friedman, and J. H. Friedman (2009): *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer.

Higuchi, Y., V. H. Nam, and T. Sonobe (2017): "Management practices, product upgrading, and enterprise survival: Evidence from randomized experiments and repeated surveys in Vietnam," *GRIPS Working Paper*.

Hirschleifer, S., D. McKenzie, R. Almeida, and C. Ridao-Cano (2016): "The Impact of Vocational Training for the Unemployed: Experimental Evidence from Turkey," *Economic Journal*, 126, 2115–2146.

Hussam, R., N. Rigol, and B. Roth (2022): "Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field," *American Economic Review*, 112, 861–98.

Iacovone, L., W. Maloney, and D. McKenzie (2022): "Improving Management with Individual and Group-Based Consulting: Results from a Randomized Experiment in Colombia," *Review of Economic Studies*, 89, 346–71.

Imbens, G. and D. Rubin (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kim, Y. R., D. Shimamoto, P. Matous, and Y. Todo (2018): "Are seminars for export promotion effective? Evidence from a randomized trial," *The World Economy*, 41, 2954–2982.

Lemoine, N. P. (2019): "Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses," *Oikos*, 128, 912–928.

Londoño, A. (2017): "Low Productivity: the Elephant in the Room in Colombia's Minimum Wage Debate," *Panam Post*.

Manski, C. (2004): "Measuring expectations," *Econometrica*, 72, 1329–1376.

McKenzie, D. (2018): "Can business owners form accurate counterfactuals? Eliciting treatment and control beliefs about their outcomes in the alternative treatment status," *Journal of Business & Economic Statistics*, 36, 714–722.

McKenzie, D. and C. Woodruff (2017): "Business Practices in Small Firms in Developing Countries," *Management Science*, 63, 2967–2981.

Meager, R. (2019): "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments," *American Economic Journal: Applied Economics*, 11, 57–91.

Otis, N. (2022): "Policy Choice and the Wisdom of Crowds," *Working Paper*.

Thorlund, K., L. Thabane, and E. J. Mills (2013): "Modelling heterogeneity variances in multiple treatment comparison meta-analysis–Are informative priors the better solution?" *BMC medical research methodology*, 13, 1–14.

Vivalt, E. (2020): "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economics Association*, 18, 3045–3089.

Table 1: Baseline Characteristics by Treatment Assignment

	Means		Std	Percentiles of Baseline Distribution				
	Control	Treatment	Dev	10th	25th	50th	75th	90th
Variables used in Stratification								
Small Firm	0.46	0.47	0.50	0	0	0	1	1
Medium Firm	0.43	0.46	0.50	0	0	0	1	1
Large Firm	0.10	0.07	0.28	0	0	0	0	0
Outlier in export data	0.09	0.10	0.29	0	0	0	0	0
Exported in last three years	0.58	0.58	0.49	0	0	1	1	1
Export Practices Index	0.36	0.38	0.22	0.09	0.18	0.36	0.55	0.64
Other Variables								
Firm located in Antioquia region	0.17	0.14	0.36	0	0	0	0	1
Firm located in Cundinamarca region	0.52	0.44	0.50	0	0	0	1	1
Firm located in Valle de Cauca region	0.07	0.15	0.31	0	0	0	0	1
Firm age (years)	20.1	20.4	13.9	5	10	18	28	40
Firm sales in 2016 (millions of pesos)	11416	11853	23564	700	1877	4596	11374	25675
Firm profits in 2016 (millions of pesos)	399	659	1492	6	36	187	505	1171
Number of employees in 2016	77	68	106	6	18	42	83	167
Business Practices Index	0.46	0.43	0.16	0.24	0.32	0.43	0.57	0.66
Administrative data on exports								
Exported at all in 2017	0.51	0.49	0.50	0	0	0.5	1	1
Winsorized number of products exported 2017	4.18	3.26	8.43	0	0	0.5	3	9.5
Winsorized number of countries exported to in 2017	2.18	1.70	3.22	0	0	0.5	2	7
Winsorized number of country-products in 2017	9.05	7.32	21.19	0	0	0.5	5	20
Number of other firms in same country-product in 2017	0.17	0.23	0.48	0	0	0	0.17	0.64
Exported a new country-product in 2017	0.44	0.34	0.49	0	0	0	1	1
Free on board export value 2017 (1000s of USD)	336	341	1170	0	0	0	171	1033
Exports per worker in 2017	7617	5922	25128	0	0	68	3806	11785
Overall Export Performance Index in 2017	0.04	-0.04	0.83	-0.72	-0.72	-0.36	0.57	1.1
Sample Size	100	100						

Table 2: Take-up rates by area

Area	Number of Firms Using	Hours Conditional on Using Area				
		Mean	SD	25th	50th	75th
Operational Productivity	71	77.8	12.9	80	80	80
Commercial	79	35	10.2	32	34.5	38
Labor Productivity	34	92.1	15.9	83	85.7	95
Energy Efficiency	19	24.3	5.3	24	26	27
Quality	32	83.5	5.4	80	80	85.5

Note: data on hours per area missing for at least four firms that took up program.

Table 3: ITT Impacts on Primary Export Outcomes

	Outcome in year:			Outcome in year:	
	2019	2020		2019	2020
Panel A: Export at all			**Panel E: Export Innovation**		
Assigned to Treatment	-0.005	-0.065	Assigned to Treatment	0.012	-0.077
	(0.043)	(0.042)		(0.056)	(0.059)
Control Mean	0.54	0.57	Control Mean	0.40	0.41
P-values:			*P-values:*	200	200
Beta = 0	0.902	0.125	Beta = 0	0.838	0.191
Beta = 0.13 (policy median)	0.002	0.000	Beta = 0.13 (policy median)	0.037	0.001
Beta = 0.06 (academic median)	0.134	0.004	Beta = 0.05 (academic median)	0.496	0.032
Beta = 0.10 (firm median)	0.016	0.000	Beta = 0.06 (firm median)	0.391	0.021
Beta = 0.09 (literature median)	0.028	0.000	Beta = 0.086 (literature median)	0.188	0.006
Panel B: Number of Products			**Panel F: Export Value**		
Assigned to Treatment	-0.142	-0.519	Assigned to Treatment	-0.375	-0.895*
	(0.377)	(0.345)		(0.487)	(0.512)
Control Mean	4.16	4.05	Control Mean	7.01	7.28
P-values:	200	200	*P-values:*		
Beta = 0	0.708	0.134	Beta = 0	0.443	0.083
Beta = 1.0 (policy median)	0.003	0.000	Beta = 0.12 (policy median)	0.311	0.049
Beta = 0.5 (academic median)	0.091	0.004	Beta = 0.12 (academic median)	0.311	0.049
Beta = 1.5 (firm median)	0.000	0.000	Beta = 0.09 (firm median)	0.342	0.056
Beta = 1.3 (literature median)	0.000	0.000	Beta = 1.37 (literature median)	0.000	0.000
Panel C: Number of Countries			**Panel G: Export Productivity**		
Assigned to Treatment	0.082	-0.179	Assigned to Treatment	-0.337	-0.671**
	(0.184)	(0.184)		(0.327)	(0.338)
Control Mean	2.33	2.37	Control Mean	4.71	4.87
P-values:	200	200	*P-values:*		
Beta = 0	0.657	0.331	Beta = 0	0.305	0.049
Beta = 0.5 (policy median)	0.024	0.000	Beta = 0.09 (policy median)	0.194	0.026
Beta = 0.5 (academic median)	0.024	0.000	Beta = 0.09 (academic median)	0.194	0.026
Beta = 0.5 (firm median)	0.024	0.000	Beta = 0.06 (firm median)	0.227	0.032
Beta = 0.66 (literature median)	0.002	0.000	Beta = 0.38 (literature median)	0.030	0.002
Panel D: Number of Product-Countries			**Panel H: Export Outcome Index**		
Assigned to Treatment	-0.027	-1.687**	Assigned to Treatment	-0.012	-0.112**
	(0.960)	(0.757)		(0.056)	(0.056)
Control Mean	10.10	9.98	Control Mean	0.03	0.08
P-values:	200	200	*P-values:*		
Beta = 0	0.978	0.027	Beta = 0	0.831	0.048
Beta = 1.5 (policy median)	0.114	0.000	Beta = 0.28 (policy median)	0.000	0.000
Beta = 1.0 (academic median)	0.287	0.001	Beta = 0.13 (academic median)	0.012	0.000
Beta = 1.5 (firm median)	0.114	0.000	Beta = 0.19 (literature median)	0.000	0.000
Beta = 4.5 (literature median)	0.476	0.002			

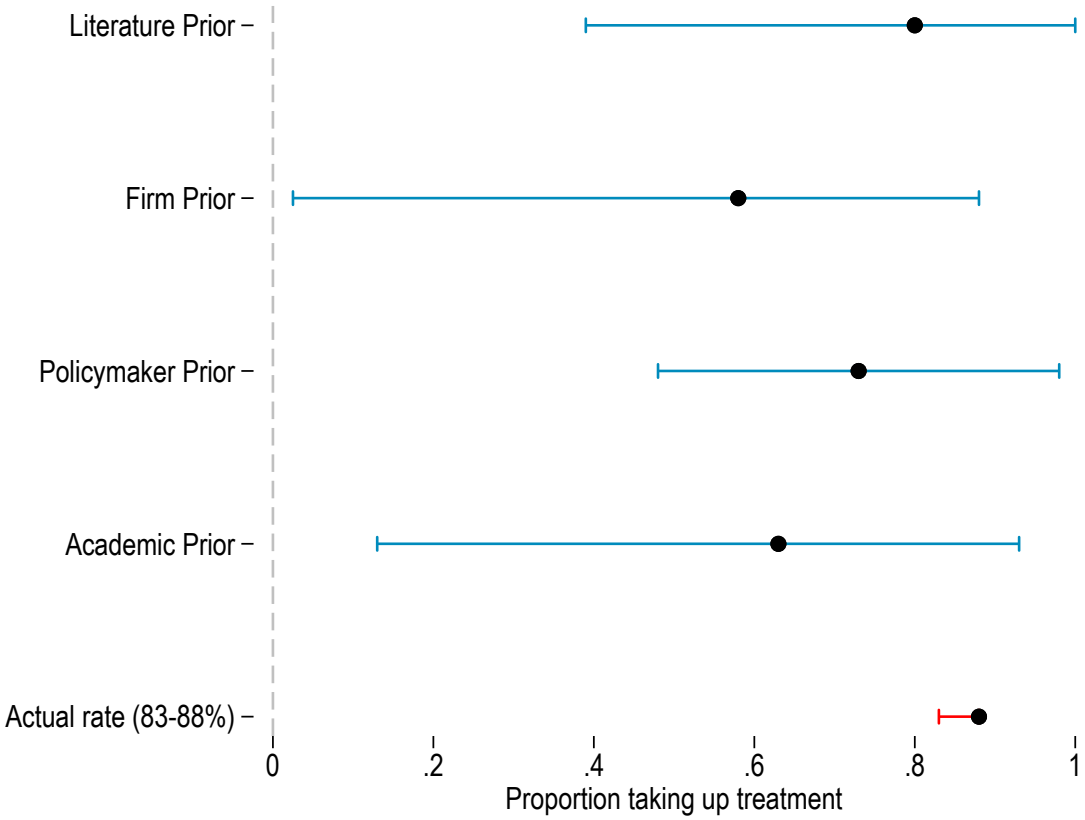
Notes: Robust standard errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels.
Sample size is 200 for all regressions. See appendix for variable definitions.

Table 4: Probability that minimum desirable effect size was achieved under various priors

Outcome	Academics	Firms	Policymakers	Literature	Default
2019					
Export At All	0.007	0.011	0.010	0.023	0.009
Number of Products	0.005	0.005	0.006	0.007	0.004
Number of Countries	0.109	0.119	0.136	0.248	0.106
Number of Product-Countries	0.181	0.186	0.304	0.154	0.117
Export Innovation	0.065	0.070	0.060	0.101	0.057
Export Value	0.676	0.693	0.870	0.804	0.197
Exports Productivity	0.170	0.215	0.567	0.767	0.106
2020					
Export At All	0	0	0	0.0003	0.0001
Number of Products	0.0001	0.0001	0.0001	0	0
Number of Countries	0.008	0.009	0.012	0.041	0.008
Number of Product-Countries	0.002	0.007	0.053	0.002	0.001
Export Innovation	0.002	0.002	0.002	0.017	0.002
Export Value	0.608	0.653	0.847	0.549	0.034
Exports Productivity	0.062	0.196	0.506	0.604	0.019

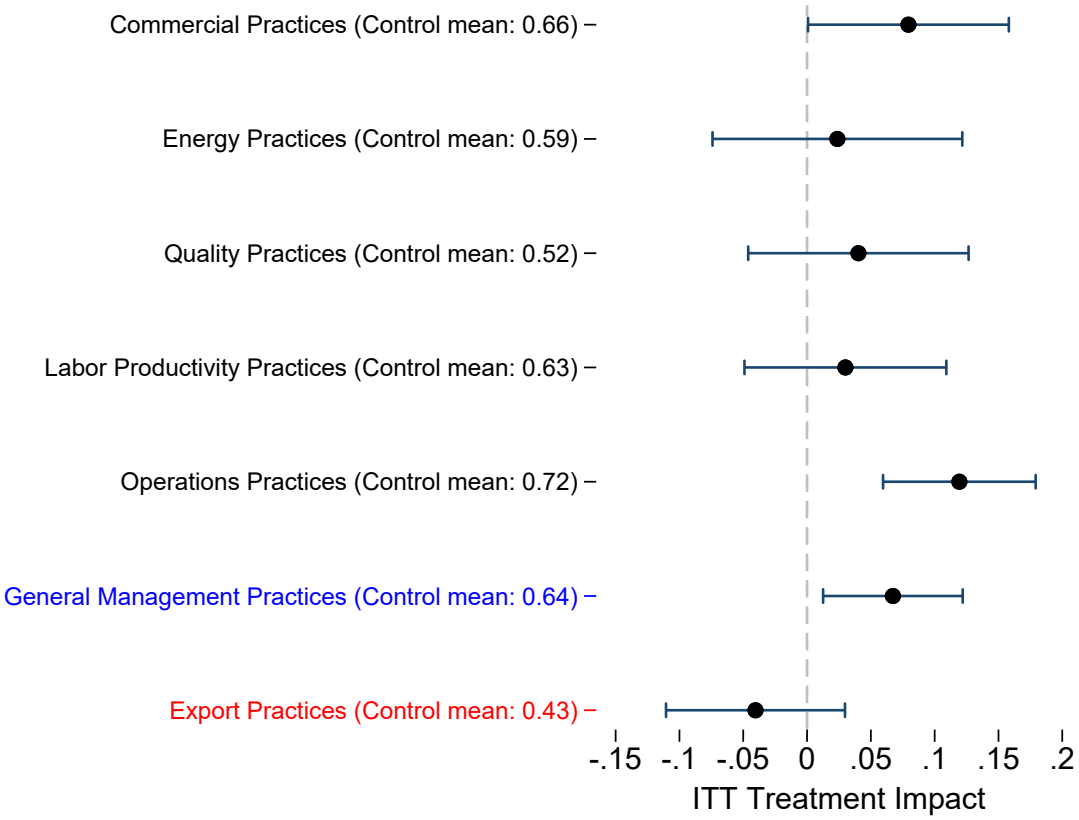
Note: Export Innovation refers to "exporting to a new product-country combination". All inference is generated by MCMC draws. Export Value and Export Productivity results reflect very little updating from priors.

Figure 1: Take-up rates compared to priors



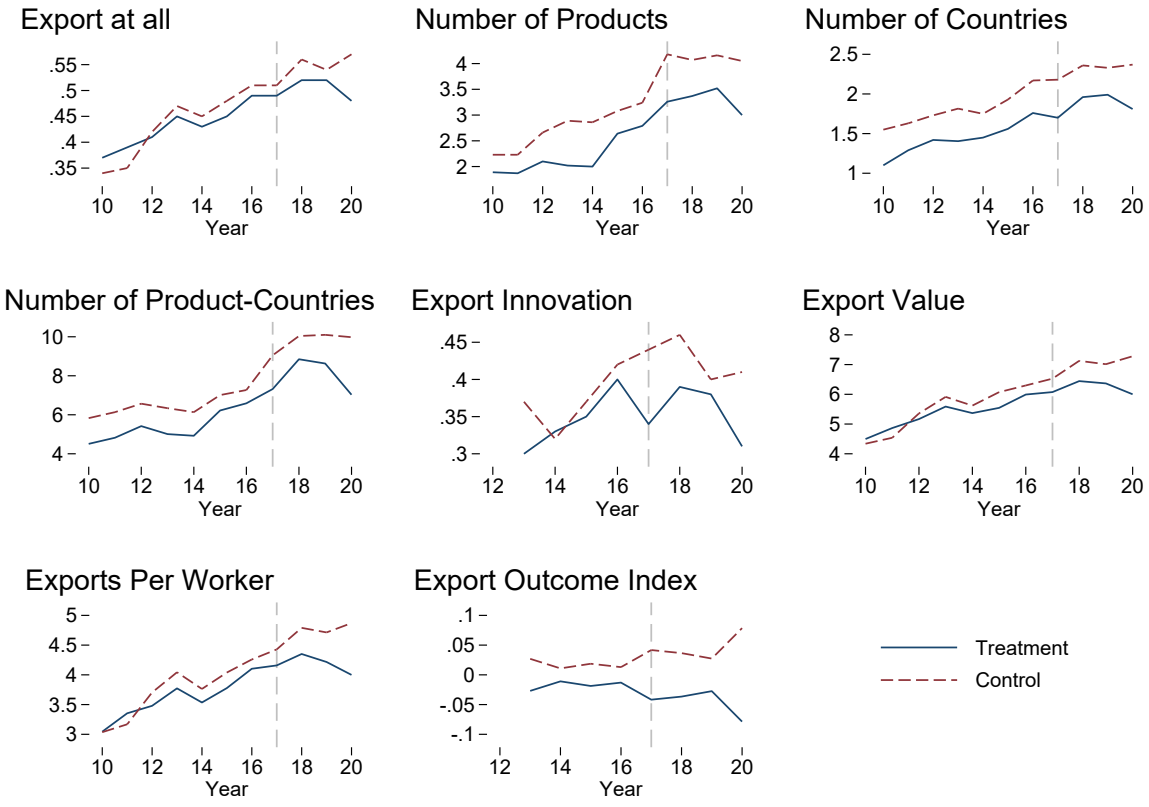
Notes: Circles show medians of elicited prior distributions, and lines show the range from the 2.5th to the 97.5th percentiles of these distributions. Actual take-up rate was 88% according to administrative data on starting the program, or 83% according to administrative data on hours of consulting received.

Figure 2: Treatment Impacts on Management and Export Practices



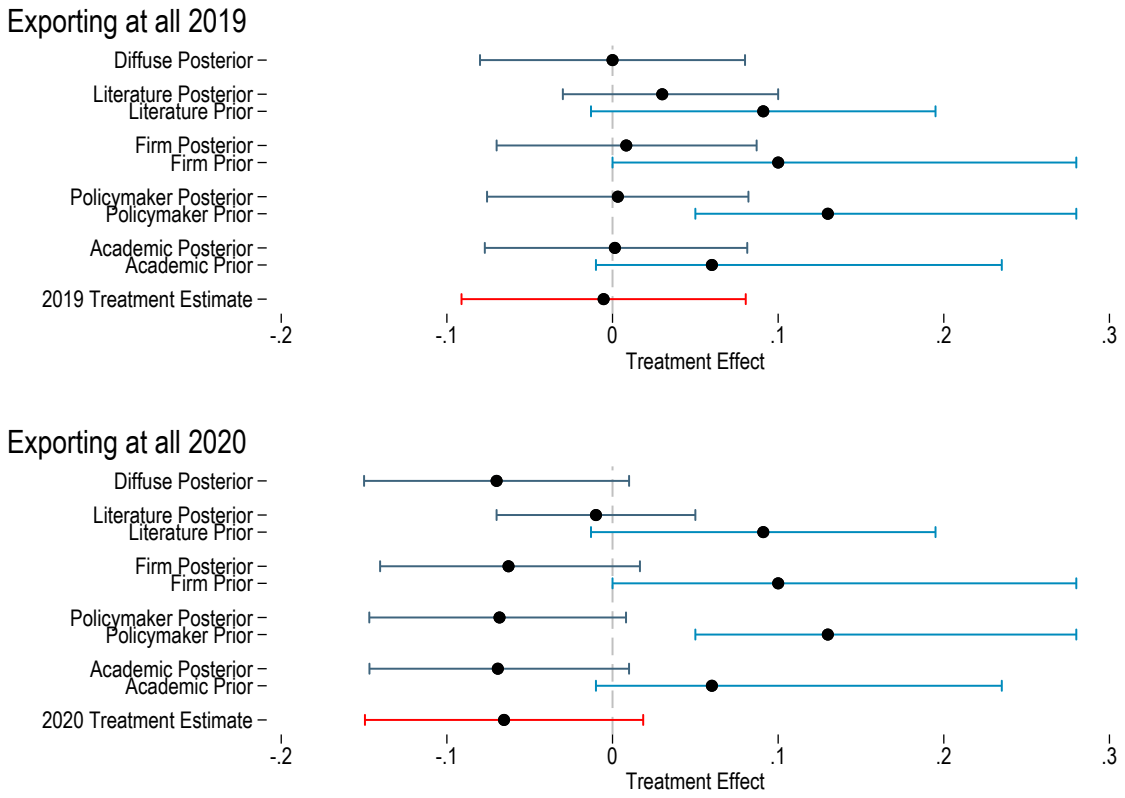
Notes: Estimated ITT treatment impacts along with 95 percent confidence intervals are shown. **Export practices** is the proportion of 15 export-specific practices implemented; **General Management Practices** is the proportion of 40 different general management practices implemented, and is comprised of five sub-indices: **Commercial Practices** (11 practices), **Energy Practices** (4 practices), **Quality Practices** (6 practices), **Labor Productivity Practices** (9 practices), and **Operations Practices** (10 practices). See Appendix C for variable definitions and impacts on sub-components.

Figure 3: Trajectory of Means of Primary Export Outcomes by Treatment Status



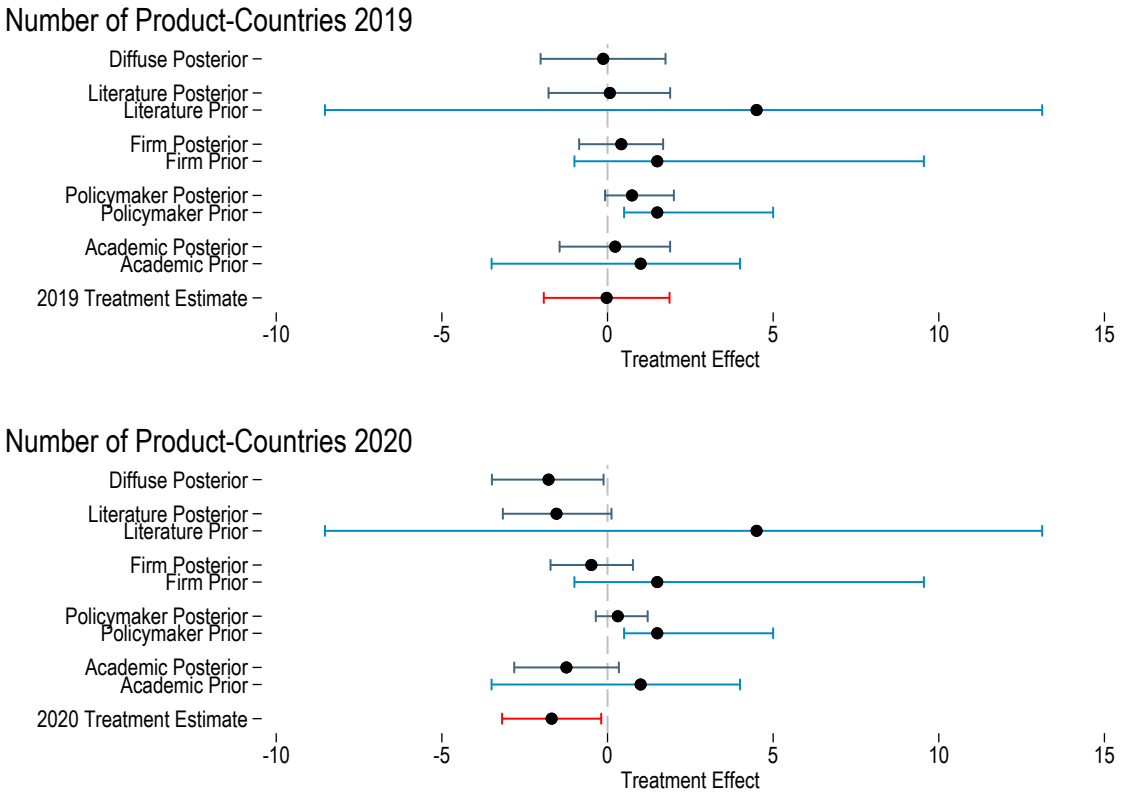
Notes: Dotted vertical line in 2017 denotes application year for program. Implementation of treatment took place in the second half of 2018 and first half of 2019. There are no significant differences between the two groups at the 5% level pre-treatment. See Appendix C for variable definitions.

Figure 4: Frequentist and Bayesian Estimation of Treatment Impacts on the Extensive Margin of Exporting at All



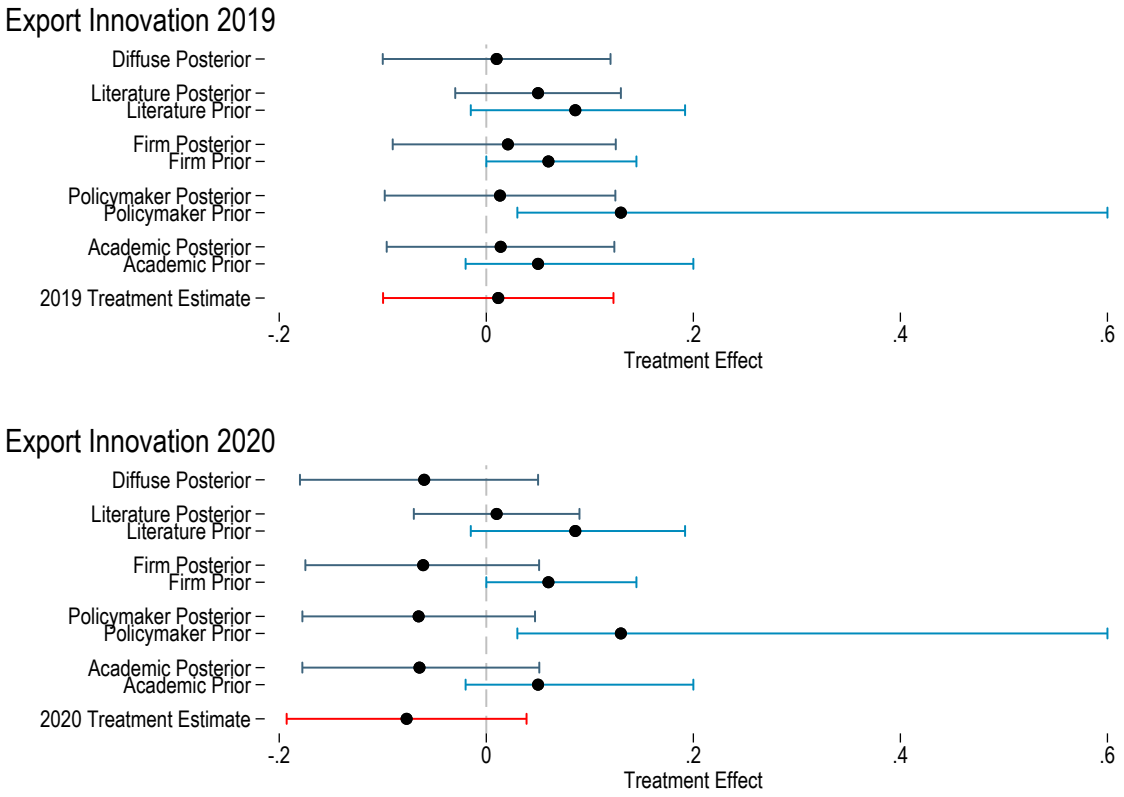
Notes: Treatment estimates and associated red line show frequentist ITT estimate and associated 95 percent confidence intervals. Treatment effect units are in terms of the change in the proportion of firms exporting. Control mean is 0.54 in 2019 and 0.57 in 2020. Light blue lines show 95 percent prior intervals elicited from academics, policymakers, and firms, and derived from the literature, with circles indicating median of prior distributions. Bayesian posteriors and dark blue lines show median and 95 percent intervals from the estimated Bayesian posterior distributions that update the associated prior with the data from the experiment. Diffuse posterior is the Bayesian posterior from using a (non-informative) diffuse prior.

Figure 5: Frequentist and Bayesian Estimation of Treatment Impacts on Export Variety



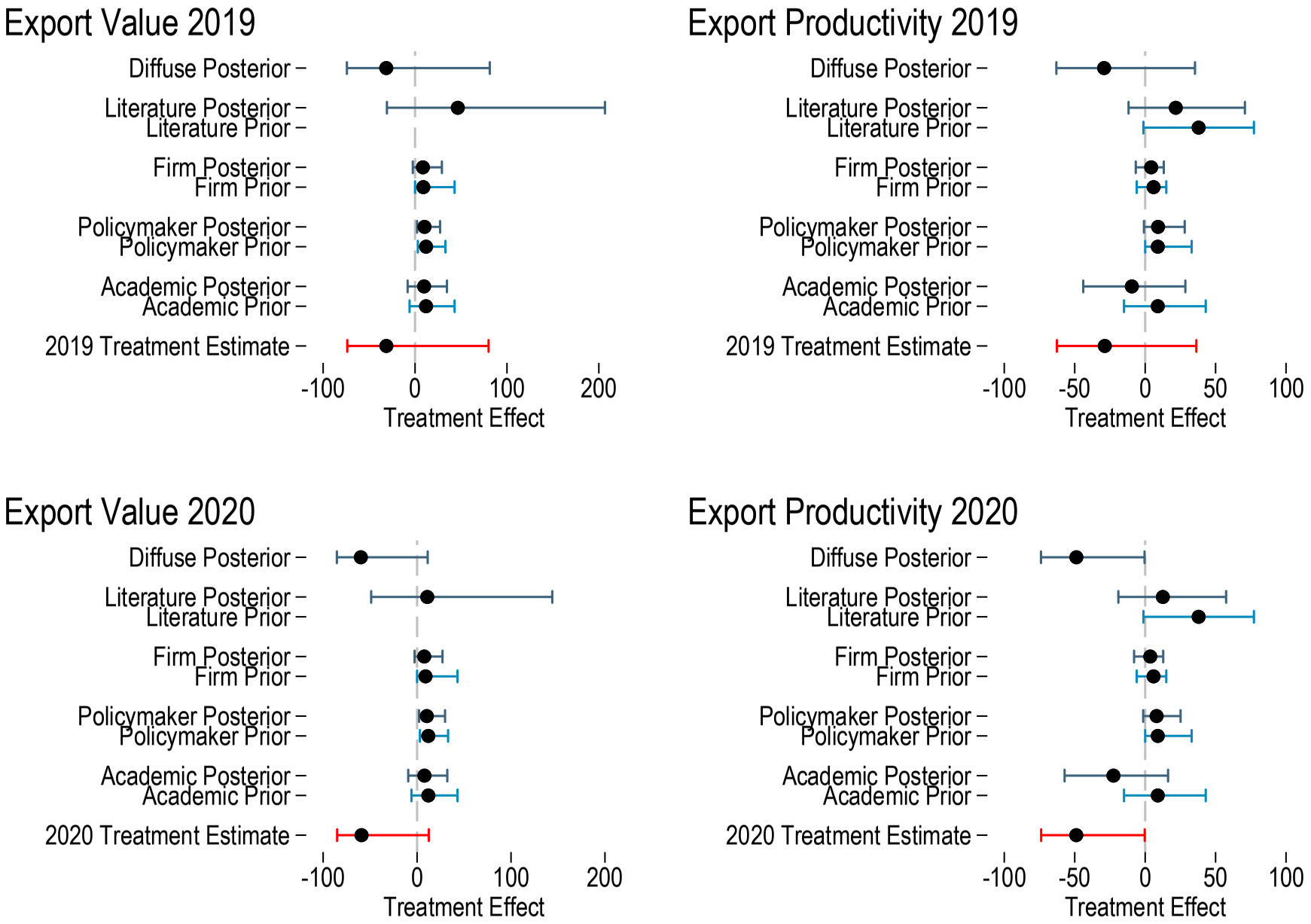
Notes: Treatment estimates and associated red line show frequentist ITT estimate and associated 95 percent confidence intervals. Number of Product-Countries is the number of distinct product-country combinations a firm exports to, and has a control mean of 10.1 in 2019 and 10.0 in 2020. Light blue lines show 95 percent prior intervals elicited from academics, policymakers, and firms, and derived from the literature, with circles indicating median of prior distributions. Bayesian posteriors and dark blue lines show median and 95 percent intervals from the estimated Bayesian posterior distributions that update the associated prior with the data from the experiment. Diffuse posterior is the Bayesian posterior from using a (non-informative) diffuse prior.

Figure 6: Frequentist and Bayesian Estimation of Treatment Impacts on Export Innovation



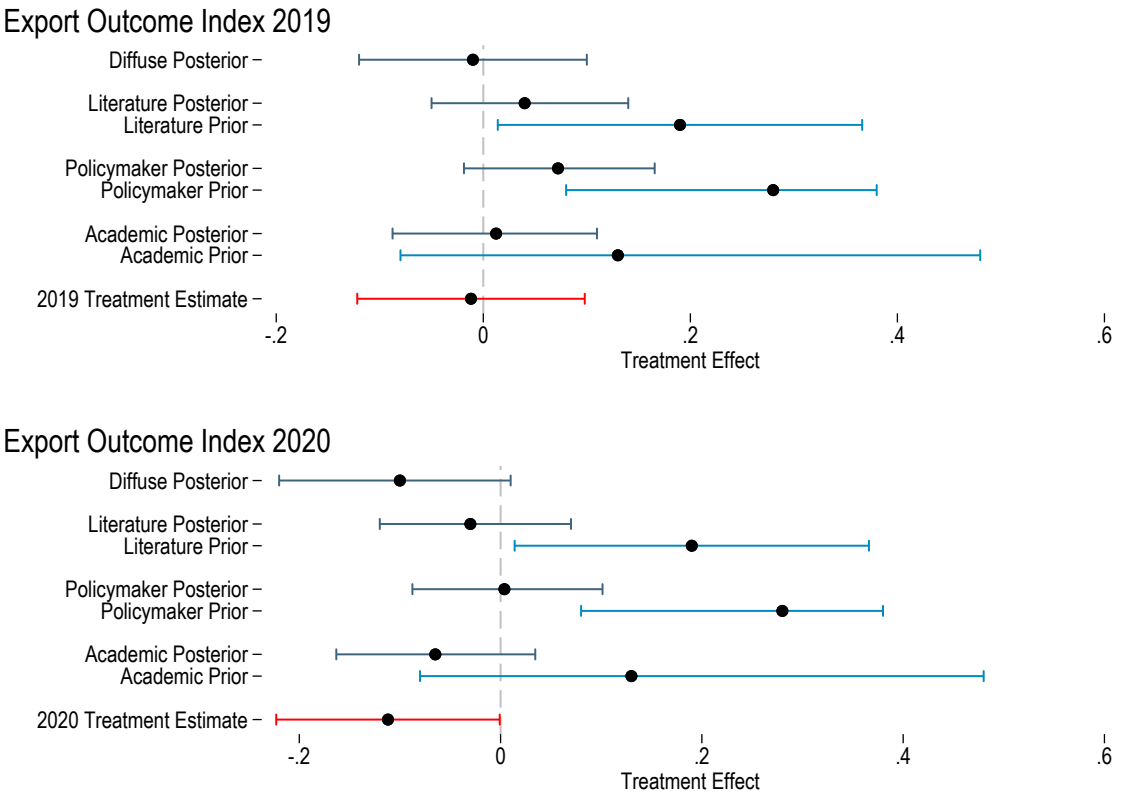
Notes: Treatment estimates and associated red line show frequentist ITT estimate and associated 95 percent confidence intervals. Export innovation is exporting a new product-country combination and has a control mean of 0.40 in 2019 and 0.41 in 2020. Light blue lines show 95 percent prior intervals elicited from academics, policymakers, and firms, and derived from the literature, with circles indicating median of prior distributions. Bayesian posteriors and dark blue lines show median and 95 percent intervals from the estimated Bayesian posterior distributions that update the associated prior with the data from the experiment. Diffuse posterior is the Bayesian posterior from using a (non-informative) diffuse prior.

Figure 7: Frequentist and Bayesian Estimation of Treatment Impacts on Export Value and Export Productivity



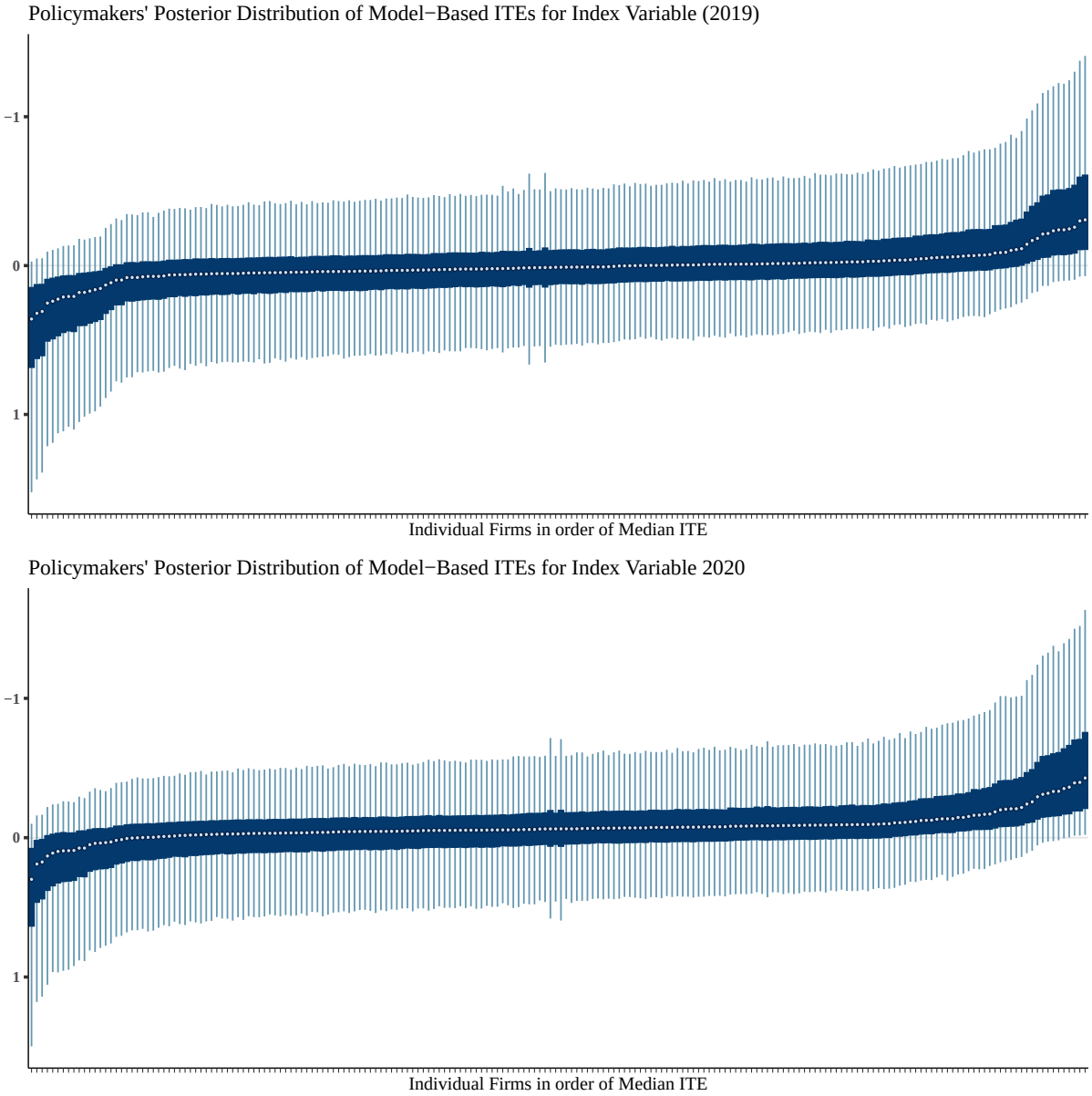
Notes: Treatment estimates and associated red line show frequentist ITT estimate and associated 95 percent confidence intervals. Treatment effects expressed in percentage terms, from treatment regressions using the inverse hyperbolic sine of the outcome. Light blue lines show 95 percent prior intervals elicited from academics, policymakers, and firms, and derived from the literature, with circles indicating median of prior distributions. Bayesian posteriors and dark blue lines show median and 95 percent intervals from the estimated Bayesian posterior distributions that update the associated prior with the data from the experiment. Diffuse posterior is the Bayesian posterior from using a (non-informative) diffuse prior. Literature prior not shown for Export Value since it is so wide it exceeds axis scale.

Figure 8: Frequentist and Bayesian Estimation of Treatment Impacts on Overall Export Performance Index



Notes: Treatment estimates and associated red line show frequentist ITT estimate and associated 95 percent confidence intervals. Overall Outcome Index is sum of standardized z-scores of seven pre-specified export outcomes. Light blue lines show 95 percent prior intervals elicited from academics, policymakers, and firms, and derived from the literature,, with circles indicating median of prior distributions. Bayesian posteriors and dark blue lines show median and 95 percent intervals from the estimated Bayesian posterior distributions that update the associated prior with the data from the experiment. Diffuse posterior is the Bayesian posterior from using a (non-informative) diffuse prior. Firms were not asked to give a prior for this outcome.

Figure 9: Bayesian Model-Based Distribution of ITEs for Policymakers



Notes: For each ITE the dark blue band shows 50% credible interval and the line is the 95% credible interval. These inferences are based on the Gaussian potential outcome distribution using elicited priors.

Online Appendices

A Timeline

Launch and Random Selection

November 2017: Launch of program

March 23, 2018: Deadline for applications

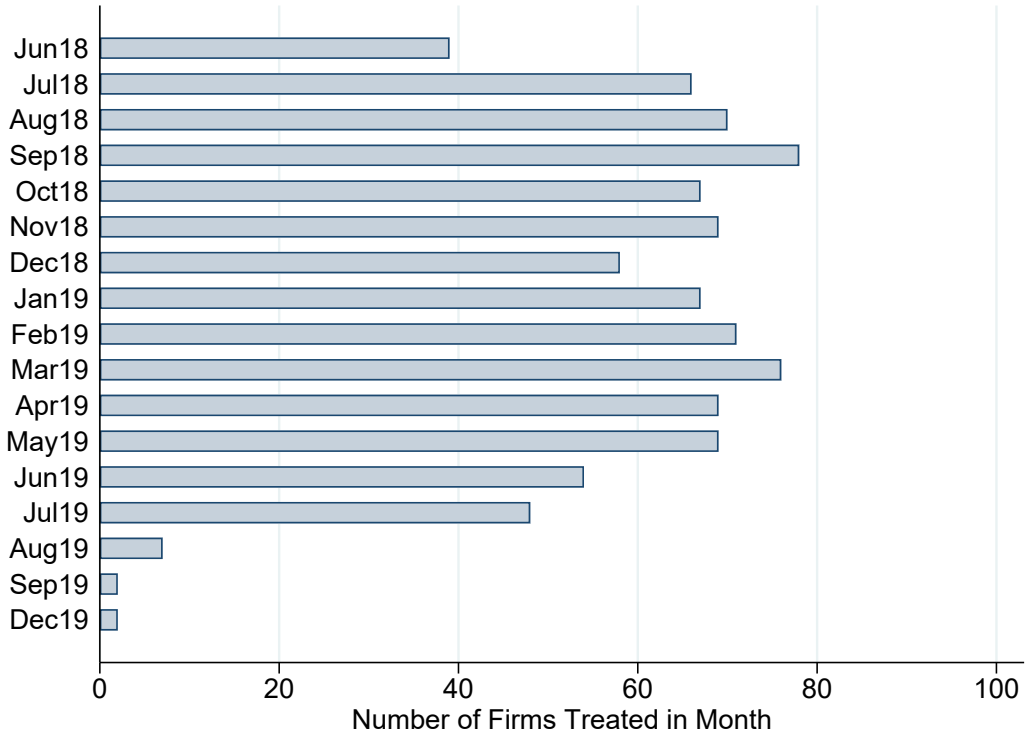
April 11, 2018: Random assignment of firms to treatment and control

Implementation

April-May 2018: Diagnostic reports done and delivered to firms in both treatment and control

The program started implementation in June 2018 and finished in December 2019. Figure A1 shows the evolution of the program's implementation by calendar month. Most firms received the intervention between July 2018 and 2019. The few firms finished after July were largely those who had selected the quality area of consulting, which had a delayed start.

Figure A1: Timing of Consulting Assistance



Source: Program administrative data.

Timing of Data Collection

June 2018-October 2018: Prior Elicitation from Policymakers, Academics, and program firms.

November 2019-May 2020: Follow-up survey of Management Practices conducted by IPA Colombia.

B More Details of the Consulting Interventions

One of the challenges with consulting interventions is understanding exactly what was done by consultants when working one on one with firms. We used qualitative fieldwork, the administrative agreements used to contract the consulting firms, interviews with the consulting companies, our own observations, and data collected from firms, and summarize key details of each component.

The **Commercial consulting** began with a lengthy diagnostic (10 hours). This included asking firms to provide data on their five main products and then identifying their star product, defined as the one they thought would have highest profitability in the future, and examining the market for this product. This was followed by an implementation phase of 23 hours, which largely focused on developing a plan for commercial strategy and exporting for the firms. For example, a cosmetics factory making many products identified sunblock as its star product. The consultants worked with them to go through their costs and client list, and suggested the firm was offering too many discounts that were incurring losses, leading the firm to cut down on promotions. A key part of this strategy was segmentation of clients, and deciding where to focus more of their sales efforts. For example, a firm making uniforms for businesses and schools was advised to focus more on the domestic market, but also to work with other Colombian agencies like Impulsa and Procolombia to develop the export market, and to take steps like setting up a webpage in English to help reach overseas buyers.

The implementing consultants for commercial management were a consortium of two companies. There were delays in starting the commercial strategy, with implementation starting in December 2018 for many firms, in part because they had not hired enough consultants. The consulting company was very focused on diagnostics, saying there is a myth that firms in Colombia are over-diagnosed, but firms we talked to complained about the amount of time and length of the diagnostic. The consulting companies had the view that before firms export, they need to know their local clients, and then that their role was to help prepare firms to work with ProColombia for any direct efforts to export. They expressed the view that they believed the majority of firms could not consistently produce the capacity to go to external markets (despite over half the sample already exporting). For some firms with more developed commercial strategies, they did

devote their implementation time towards more specific export-oriented activities, such as studying how to enter the Ecuadorian market, or offering advice on talking with potential clients in the United States.

The **Operational Productivity consulting** appears to have worked primarily on efforts to standardize processes, reduce bottlenecks, and cut stoppages and dead time. One of the key tools used was value-stream mapping (VSM), a lean manufacturing method that maps the flow of materials through production and the time taken for each step. For example, in one factory we visited, the consultant had gone through with the firm to outline step-by-step in the production process the amount of time, manpower, and moving distance involved in each step. They then suggested how improvements could be made by reducing the number of times an item gets moved back and forward in production.

Some firms noted that these approaches had helped them with production efficiency. However, several challenges arose. A first one is that the methodology may not be well suited for all firms, especially multi-product firms making a wide variety of products and where the production steps may differ across products. One textile firm we visited noted that after going through the VSM process, they concluded that it was not working that well for their purposes, since the production flow changes a lot from clothing item to item. A second potential drawback is that these activities often took place before the firm worked on its commercial strategy, so that the focus was on lowering cost and improving efficiency for a particular existing product, but not, for example, "we should focus more on developing a higher quality brand if we want to export to the US/Europe, using organic products and better packaging, and then design the production process around this". This is a general challenge for the program, since different consulting firms are contracted to work on the different types of management improvement.

The **Labor Productivity consulting** appears to have focused largely on helping firms retain talent and improving worker morale. It was recommended that firms do an employee climate survey, to get a sense of employee views, attitudes towards the organization, and issues faced. A common issue identified was that production-level workers did not feel that involved in decision-making, and so in several firms a recommendation was made to start each day with a short line-level meeting where workers could talk about production indicators and any problems in operations. Other firms introduced worker suggestion boxes or other ways for workers to give feedback. Firms were also advised to introduce non-monetary recognition programs, and consider other morale-boosting activities like social events.

Some of the firms we spoke to noted that the consultants lacked sector-specific knowledge about specific labor force issues in their industries, and that a challenge was that some of the ways to motivate labor productivity depended on first having sorted out operational productivity indicators first. The sequencing and coordination with the operational

consulting was sometimes an issue for this.

The **Quality consultants** were the last to start due to contracting issues. They saw their role as supporting companies to implement quality standards that could help them overcome technical barriers to accessing markets, especially those covered under trade agreements. For example, a firm making chocolate and other cocoa products said they received help to improve the quality standards under which they operate which they hoped would help them to meet export requirements.

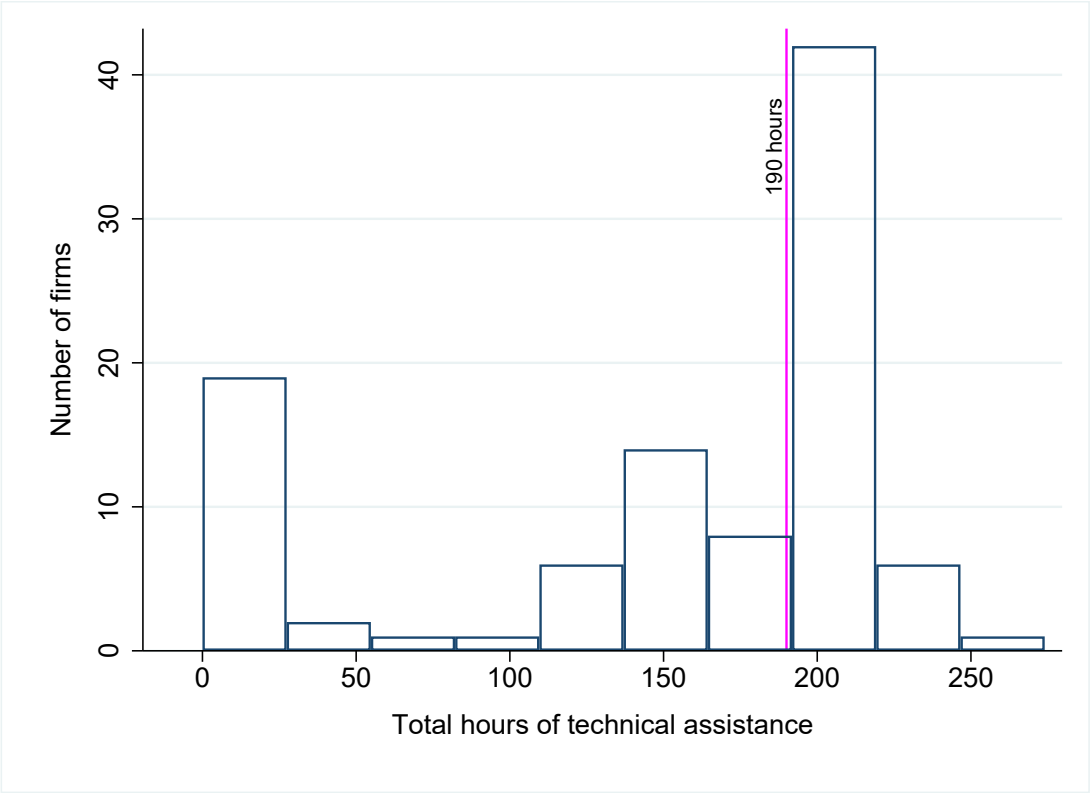
The **Energy consultants** worked on helping the firms identify opportunities for energy savings. For example, a firm making flour and cereals introduced new LED lighting to help reduce energy costs. A company making essential oils made a change to reduce the use of diesel-powered electricity by introducing solar panels.

Overall, our qualitative interviews found firms happiest with the productivity area consultants, who they found to be well organized and with clear ideas around implementing lean methodology. They were least happy with the commercial strategy consultants. Several issues ran across the different areas of consulting. A first was the lack of coordination amongst the consultants in the different areas. The ordering of which area of consulting they received first was haphazard, with no strategic vision. A common theme across most areas was that the diagnostic and action plan phases took a lot longer than managers would have liked, which was particularly difficult for smaller firms in which a manager or director may work on multiple functions. Another issue was that of consultant turnover. It appears that some of the consulting firms had to hire additional staff to work on this assignment, and there was a lack of continuity from one consultant to the next when a consultant quit.

Both the treatment and control firms were also invited to a trade fair (*macrorrueda de negocios*) in April 2019, organized by a separate agency, ProColombia, and designed as a business matchmaking forum to enable firms to obtain meetings with international buyers. However, in practice, the heterogeneity of firms made it difficult to organize buyers for each sector, and coupled with interagency logistical frictions, meant that this was poorly attended, with only 43 of the 200 firms (26 treatment and 17 control) attending. For example, one firm reported that it had received an email from Colombia Productiva inviting them to this event, and then a cancelation of this invitation a couple of months later because there were insufficient potential clients to make it worth them attending. Another firm which did attend reported that it did not find it very useful, since the few potential clients that were there were looking for the lowest price producer and they could not compete on price alone.

Figure B1 shows the distribution of total hours of consulting received by firms in the program.

Figure B1: Distribution of Total Hours of Consulting (All Firms)



Source: Program administrative data. Firms with no hours of consulting recorded are shown as receiving zero hours..

C Data and Measurement Details

Our primary outcomes use annual data on exports from 2010 to 2020 provided by the National Directorate of Taxes and Customs (DIAN) and supplied to us by the Colombian National Planning Department (DNP). Our outcomes are defined as follows:

1. **Extensive margin: Export at all in the past year**: This is a binary variable, defined as one if the firm exports directly at all in the year, and zero otherwise.

2. **Number of Distinct Products Exported in the past year**: The number of different product categories exported in the past year, using the 6-digit product classification in the harmonized system for the Andean Community. This is coded as zero for firms that do not export, and is winsorized at the 99th percentile.

3. **Number of Different Countries Exported to in the past year**. The number of different countries the firm exported to in the past year, coded as zero for firms that do not export, and winsorized at the 99th percentile.

4. **Number of Distinct Product-Country Combinations Exported in the past year**: This counts the number of product-country combinations a firm exported to

in the past year, coded as zero for firms that do not export, and winsorized at the 99th percentile.

5. **Export innovation (new product-country combination)**: This is a binary variable coded as one if the firm exported to a product-country pair that they had not exported to at all in the past three years, and zero otherwise. Coded as zero for firms that do not export.

6. **Inverse Hyperbolic Sine of Total Export Value in the past year**. This takes the inverse hyperbolic sine transformation of total exports (measured in US dollars), and includes exports coded as zero for firms that do not export.

7. **Inverse Hyperbolic Since of Export Labor Productivity**: This is the inverse hyperbolic sine of the ratio of total export value in the past year (measured in US dollars) to the average number of workers used in the past year (obtained from the PILA database, which has monthly data on formal workers). This is coded as zero for firms that do not export, and winsorized at the 99th percentile.

8. **Standardized export outcomes index**: This index is calculated as the average of the normalized z-scores of primary export outcomes 1 through 7, where each z-score is defined by subtracting the mean and dividing by the standard deviation of the respective outcome.

Our measures of general management and export-specific practices come from survey data collected by Innovations for Poverty Action. Our outcomes are defined as follows:

1. **Export Practices**: The proportion of 15 export-specific practices being used by the firm, coded as zero for firms that close. The 15 practices measured are whether the firm has done any of the following in the past 12 months: received quality certification on export products; planned production with time for external markets; planned resources or training needed for export processes; participated in missions of public agency offerings to learn export quality standards; traveled to external markets to understand consumers there; participated in a trade fair or other industry exhibition for foreign visitors; developed direct marketing strategies to reach international customers; implemented new productive techniques to meet international quality standards; developed a new product with export potential; received distributor information through advertising abroad; received distributor information through an export promotion agency; received distributor information through trade fairs; received distributor information through commercial missions; evaluated the export potential of the company's products; examined information about customer preferences and consumption patterns in external markets. Figure C1 shows the treatment impacts on each individual component.

2. **General Management Practices**: The proportion of 40 different management practices being used by the firm, coded as zero for firms that close. These are sub-divided into five areas and component indices:

   (a) **Commercial Practices**: The proportion of the following 11 practices implemented in the past 12 months: a marketing and sales plan aligned with business strategy; a market segmentation approach with customers; market research identifying current and potential customers in the domestic market; market research identifying potential retailers and wholesalers in the domestic market; use a customer relationship management (CRM) system to record customer information; had meetings with customers to get feedback on products and services; promoted products and services by advertising and brand improvements; clearly defined the role of sales staff in hiring process; provided sales training courses for sales staff; created incentive schemes for sales staff; measured key performance indicators (KPIs) for commercial practices. Figure C2 shows treatment impacts on each individual component.

   (b) **Operational Productivity Practices**: The proportion of the following 10 practices implemented in the past 12 months: communicated strategic goals to the production plant leader and team; evaluated the plant lead and team according to strategic goals; identified and managed bottlenecks in the company's capacity; identified and described production processes using Value Stream Mapping (VSM); developed the 5S methodology in the production plant; developed continuous improvement methods (Kaizen, Takt time) in the production process; standardized the work sequence by breaking down every component of the production process; implemented preventive and corrective measures; measured key performance indicators (KPIs) on plant performance; developed improvement programs for KPIs. Figure C2 shows treatment impacts on each individual component.

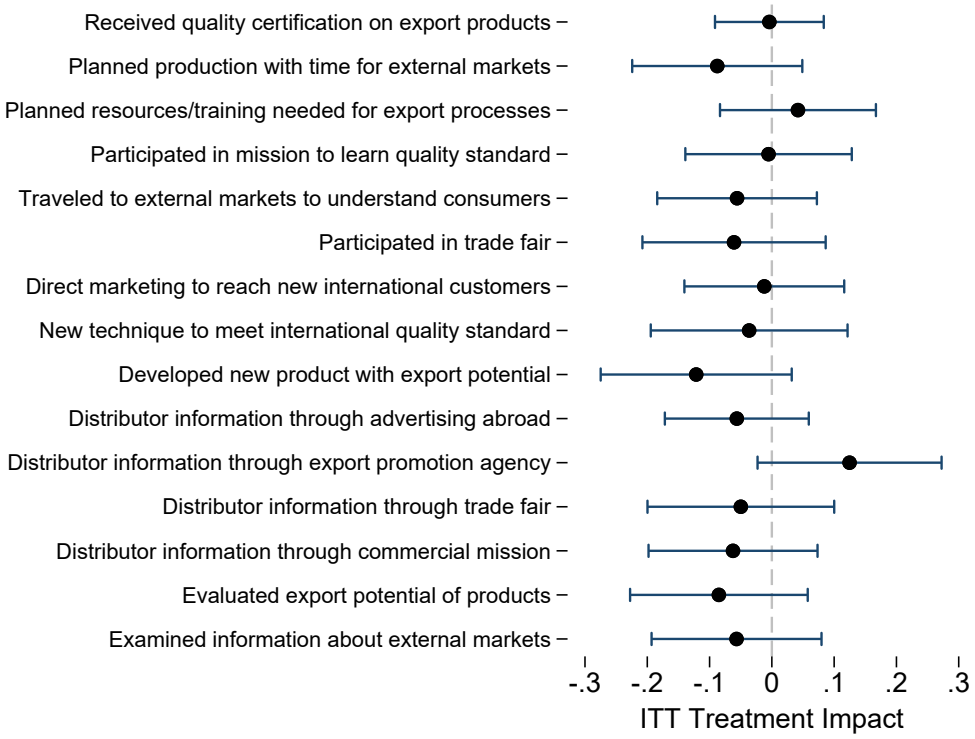
   (c) **Labor Productivity Practice**s: The proportion of the following 9 practices implemented in the last 12 months: defined a talent management plan aligned with the strategic plan; developed communication practices among the different processes of the company; promoted a culture of measuring work environment; improved workspaces for employees (ergonomics, illumination); developed programs leading to changes in organizational climate, such as communication workshops; administered satisfaction surveys with employees; developed workshops and trainings regarding absenteeism; created a rewards system or recognition program based on employee performance; measured KPIs for labor productivity.

   (d) **Quality Practices**: The proportion of the following 6 practices implemented

in the last 12 months: Pursued certification in the quality of a product; pursued certification in the quality of a process; advanced in the implementation of a certain code in the pursued quality certification; defined formal indicators to measure quality; advanced in the development of quality activities in the working plan; documented any of the production processes of the company.

(e) **Energy Practices**: The proportion of the following 4 practices implemented in the last 12 months: established energy efficiency indicators to identify energy savings opportunities; created maintenance protocols for the equipment used in production; installed or improved illumination such as changing to LED; measured KPIs for energy.

Figure C1: Impacts on Individual Components of Export Practices Index



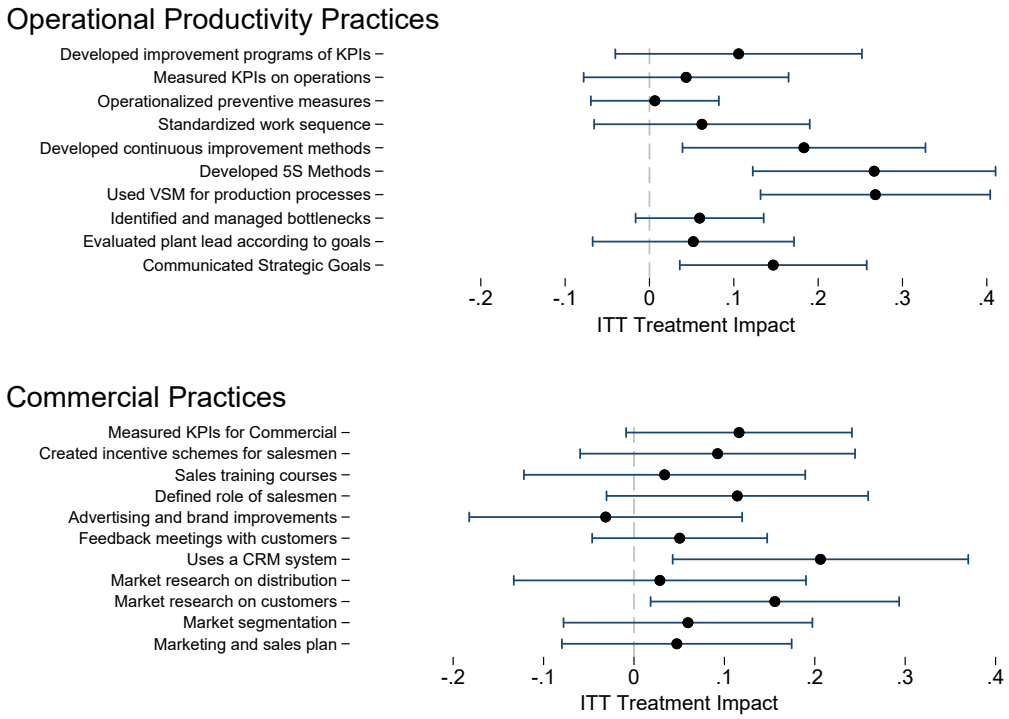
Notes: Estimated ITT treatment impacts along with 95 percent confidence intervals are shown.

D Examples of Prior Elicitation

Example of language used in explaining ITT

Beliefs About the Impact of the Program on Those Offered the Full Program We are now going to ask you for your beliefs for the likely impact of the program for the firms that are offered the full intervention, compared to firms offered just the diagnostic phase and trade

Figure C2: Impacts on Individual Components of Operational Productivity and Commercial Practices Indices



Notes: Estimated ITT treatment impacts along with 95 percent confidence intervals are shown.

fair. In technical terms, we want to know your beliefs about the intention-to-treat effect (ITT). This is the effect of BEING OFFERED the full program, regardless of whether or not a firm takes up the program. It is simply the DIFFERENCE IN MEAN OUTCOMES for the 100 firms that get the full intervention compared to the 100 firms that just get the diagnostic and trade fair. For example, suppose we look at the percentage growth in exports. We might imagine the 100 firms in the full intervention group can be broken down into:

- 10 firms that decide not to pay for the program, and see no benefit: 0% export growth
- 10 firms that take up the program, but don't see any benefit from it: 0% export growth
- 40 firms that take up the program and see small improvements, perhaps a 10% increase in exports: 10% export growth
- And 40 firms that take up the program and see large gains of 50% export growth each

Then the MEAN export growth for the full intervention group is 0.1*0% + 0.1*0% + 0.4*10% + 0.4*50% = 24% And suppose those firms who just get offered the diagnostic and trade fair have an average export growth of 5% Then the intention-to-treat effect is the difference in these two groups, which is 24-5% = +19%. So for the set of questions which follow, we are interested in learning what you think will be the difference in the average outcome for the 100 firms OFFERED the full program, compared to the 100 firms that get offered only the basic program of a diagnostic and trade fair.

Figure D1: Example of Grid Used for Eliciting Take-up Prior

Please allocate 20 stones into the different ranges below, according to how likely you think it is that the number of firms out of 100 in the program that choose to pay their share of the cost and receive the full intervention lies in each range. For example, if you think there is a 10 percent chance that between 16% and 20% of firms will end up taking up the program, put 2 stones (write 2) in the cell under 6 to 10, and allocate your other 18 stones according to what else you think is likely.

0 to 5	6 to 10	11 to 15	16 to 20	21 to 25
26 to 30	31 to 35	36 to 40	41 to 45	46 to 50
51 to 55	56 to 60	61 to 65	66 to 70	71 to 75
76 to 80	81 to 85	86 to 90	91 to 95	96 to 100

Figure D2: Example of Grid for Eliciting Priors on ITT of Export Extensive Margin

The 190 hours of technical assistance will begin in the second half of 2018. 49 percent of the firms in the Benefits 2 (full intervention group) exported in 2017.

We want to know how much you think this will CHANGE for the group getting offered the full intervention compared to getting offered just the diagnostic and trade fair, over the first 12 months since firms start their implementation. For example, if you think there is a 15 percent chance the intervention will increase the percent of firms exporting by 8 percentage points, put 3 stones in the box under 8, and allocate your remaining stones according to what else you think is likely.

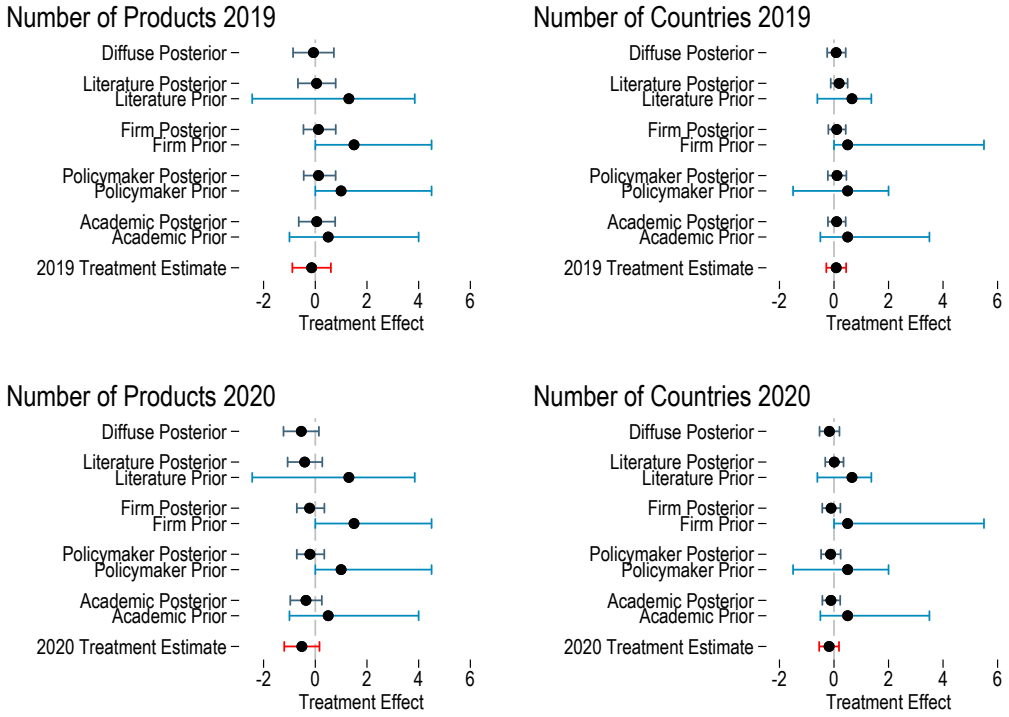
Where a single number is written (e.g. 7, you should think of it as the interval [7.0 to 7.99]

BELIEFS ABOUT THE IMPACT ON THE PERCENT OF FIRMS THAT EXPORT

-51 or less	-31 to -50	-21 to -30	-16 to -20	-11 to -15
-10	-9	-8	-7	-6
-5	-4	-3	-2	-1
0	1	2	3	4
5	6	7	8	9
10	11	12	13	14 to 15
16 to 18	19 to 21	22 to 25	26 to 30	31 to 35
36 to 40	41 to 45	46 to 50	51 to 60	>60

E Impacts on Other Outcomes

Figure E1: Impacts on Number of Products and Countries Exported



Notes: Treatment estimates and associated red line show frequentist ITT estimate and associated 95 percent confidence intervals. Treatment effect units are in terms of the change in the number of products or countries. Control Mean is 4.16 products and 2.33 countries in 2019, and 4.05 products and 2.37 countries in 2020. Light blue lines show 95 percent prior intervals elicited from academics, policymakers, and firms, and derived from the literature, with circles indicating median of prior distributions. Bayesian posteriors and dark blue lines show median and 95 percent intervals from the estimated Bayesian posterior distributions that update the associated prior with the data from the experiment. Diffuse posterior is the Bayesian posterior from using a (non-informative) diffuse prior..

We use additional data sources to examine impacts on secondary outcomes. We did not collect priors on any of these outcomes, so present frequentist treatment impacts only.

Impacts on Survival and Employment Using the PILA

Our employment data come from the *Planilla Integrada de Liquidación de Aportes* (PILA) (Unified Register of Contributions), which is the national information system used by firms to file the mandatory contributions to health, pensions, and disability insurance paid for workers. Of the 200 firms that applied for the program, 198 were able to be matched to the PILA, 195 have data in early 2018 (pre-intervention), and 185 are found to be reporting data in December 2020.

Panel A of Table E2 examines the impact of the program on firm survival. Columns 1 and 2 define survival as showing up in the PILA, and look at survival until the end of 2019 and end of 2020 respectively. We see 4 percent of control firms have died by the end

of 2019, and 7 percent by the end of 2020. Firms assigned to treatment are more likely to survive by 3 percentage points in 2019, and 5 percentage points in 2020. Firms may not appear in the PILA because they have closed, or because of reporting issues. For example, a couple of the firms appear to be registered only under the owner's name as a natural person, rather than as companies. We therefore consider a second definition of survival, which involves cross-checking data from formal employment reported in the PILA with data on the firm from the export database, from the RUES (Mercantile Registry), and from IPA calling and visiting firms to verify their status. We believe this is the most accurate measure of firm survival as a result. Column 3 looks at impacts on this measure of survival for the full 200 firms, and again finds treated firms are 5.1 percentage points more likely to survive, with this effect significant at the 10 percent level.

The firms that closed were considerably smaller to begin with on average than those that survived. For the control group, mean (median) January 2018 employment was 87 (48) for the survivors, compared to 51 (21) for firms that closed; for the treatment group it was 73 (43) for the survivors, compared to 23 (11) for firms that closed. Moreover, the firms that died were much less likely to be engaged in exporting: 0 out of the control firms that died had exported in the past three years prior to applying for the program compared to 64 percent of the control firms that survived, and only 1 of the treated firms that died had exported in the three years prior to applying for the program. Since we code exports as zero for firms that are closed, closure effectively results in replacing 0 exports for an open firm with 0 exports for a closed firm.

Panel B of Table E2 provides treatment estimates on monthly firm employment using the PILA. Column 1 and 2 show an impact on the level of employment of 0.49 workers relative to a control mean of 81 workers in 2019, and of 0.44 workers relative to a control mean of 79 workers. This effect is small in magnitude (0.6% of the control mean), and not statistically significant. Column 3 restricts to the sample of surviving firms, and again finds a small impact. The 95 percent confidence interval for the treatment impact in 2020 is (-3.5 workers, +4.4 workers), which includes changes of up to 5 percent in employment. Employment changes may take longer to manifest, but at least in the first two years, the program did not result in large changes in employment.

Impact on Sales, Profits and Labor Productivity Using the RUES After our pre-registration, we became aware of an additional source of administrative data on firms, which is the *Registro Único Empresarial* (RUES), which contains data from firms' annual renewal in the *Registro Mercantil* (Mercantile Registry).[13] Colombia Productiva provided

[13]Our pre-analysis plan also said we would attempt to link the firms to the Encuesta Anual Manufacturera (Annual Manufacturing Survey, or EAM). These data are released with a considerable lag, and can only be accessed in a data lab in Colombia, with the pandemic making access to the lab more limited. We had a consultant work with these data, with only data up to the end of 2019 available. 181 firms were matched in the 2017 data, and 182 firms in the 2018 and 2019 data. The point estimates suggest a

us these data for 2018, 2019 and 2020 only, since a change in reporting precluded earlier years. 199 out of the 200 firms were able to be matched to this registry, with the only exception a firm that went bankrupt between applying and the program starting. The registry has fewer firms reporting sales and employment in 2019 than in 2020, which we believe may reflect the start of the pandemic in 2020 stopping some firms from doing their annual reporting for the 2019 year. Since we lack pre-program data, we control for randomization strata, and then use post-double selection lasso to select controls from our set of baseline variables.

We use these data in panels C, D and E of Table E2 to measure impacts on sales, profits, and sales per worker (a measure of labor productivity). We show results for both 2019 and 2020, in both levels and inverse hyperbolic sines, given the heterogeneity in firm size. When we measure impacts on levels, the coefficients are all negative, and more so in 2020 than 2019. The magnitudes relative to the control mean in percentage terms are for a 13 percent reduction in sales, 30 percent reduction in profits, and 18 percent reduction in sales per worker. However, the standard errors are large, and only the impact on sales in 2020 is statistically significant. The inverse hyperbolic sine transformation places more weight on smaller firms when calculating the average, and on the extensive margin of surviving (not having zero sales or profits). This results in a more mixed pattern of some positive and some negative coefficients, but with large standard errors, and none of the results are statistically significant. While the results do not preclude that the improvement in general management practices increased firm productivity, there is no strong evidence of this having occurred to date.

fall in total production and in total factor productivity in 2019, but standard errors are large and none of the estimates are statistically significant.

Table E1: Robustness of Export Results to Dropping Outlier Strata

	Full Sample		No Outlier Sample	
	2019	2020	2019	2020
Panel A: Export at all				
Assigned to Treatment	-0.005	-0.065	-0.007	-0.073
	(0.043)	(0.042)	(0.048)	(0.047)
Control Mean	0.54	0.57	0.51	0.54
Panel B: Number of Products				
Assigned to Treatment	-0.142	-0.519	-0.021	-0.593*
	(0.377)	(0.345)	(0.299)	(0.336)
Control Mean	4.16	4.05	3.12	3.08
Panel C: Number of Countries				
Assigned to Treatment	0.082	-0.179	0.002	-0.159
	(0.184)	(0.184)	(0.200)	(0.203)
Control Mean	2.33	2.37	1.93	1.97
Panel D: Number of Product-Countries				
Assigned to Treatment	-0.027	-1.687**	0.256	-1.084*
	(0.960)	(0.757)	(0.549)	(0.595)
Control Mean	10.10	9.98	7.48	7.36
Panel E: Export Innovation				
Assigned to Treatment	0.012	-0.077	0.019	-0.053
	(0.056)	(0.059)	(0.061)	(0.061)
Control Mean	0.40	0.41	0.36	0.36
Panel F: Export Value				
Assigned to Treatment	-0.375	-0.895*	-0.394	-0.970*
	(0.487)	(0.512)	(0.535)	(0.562)
Control Mean	7.01	7.28	6.44	6.74
Panel G: Export Productivity				
Assigned to Treatment	-0.337	-0.671**	-0.369	-0.740**
	(0.327)	(0.338)	(0.360)	(0.370)
Control Mean	4.71	4.87	4.35	4.53
Panel H: Export Outcome Index				
Assigned to Treatment	-0.012	-0.112**	-0.012	-0.105*
	(0.056)	(0.056)	(0.060)	(0.062)
Control Mean	0.03	0.08	-0.06	-0.01
Sample Size	200	200	181	181

Notes:
Regressions control for five pre-treatment annual lags of dependent variable and for randomization strata. Robust standard errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively.
No outlier sample drops a strata with high baseline exporting.

Table E2: Impacts on Secondary Outcomes

	(1)	(2)	(3)	(4)
Panel A: Firm Survival				
	Dec 2019	Dec 2020	Dec 2020	
Assigned to Treatment	0.032	0.053	0.051*	
	(0.026)	(0.034)	(0.030)	
Control Mean	.96	.93	.93	
Sample Size	195	195	200	
Sample	PILA	PILA	PILA+cross-check	
Panel B: Formal Employment				
	2018:5-2019:12	2018:5-2020:12	2018:5-2020:12	
Assigned to Treatment	0.487	0.440	0.211	
	(1.627)	(2.003)	(2.037)	
Control Mean	81	79	85	
Sample Size	4000	6400	6067	
Sample	PILA	PILA	PILA (survivors only)	
Panel C: Firm Sales				
	2019 Level	2019 I.H.S.	2020 Levels	2020 I.H.S.
Assigned to Treatment	-307	0.467	-1817**	0.319
	(615)	(0.493)	(767)	(0.392)
Control Mean	11138	22.2	14346	22.5
Sample Size	173	173	198	198
Panel D: Firm Profits				
	2019 Level	2019 I.H.S.	2020 Levels	2020 I.H.S.
Assigned to Treatment	-246	-0.335	-338	-0.875
	(207)	(1.311)	(207)	(1.357)
Control Mean	1020	16.64	1118	17.06
Sample Size	200	200	200	200
Panel E: Labor Productivity (Sales per Worker)				
	2019 Level	2019 I.H.S.	2020 Levels	2020 I.H.S.
Assigned to Treatment	-2.7	0.376	-39.3	-0.135
	(22.2)	(0.263)	(31.1)	(0.100)
Control Mean	170	19.04	223	19.52
Sample Size	162	162	190	190

Notes:

Notes: Robust standard errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels.

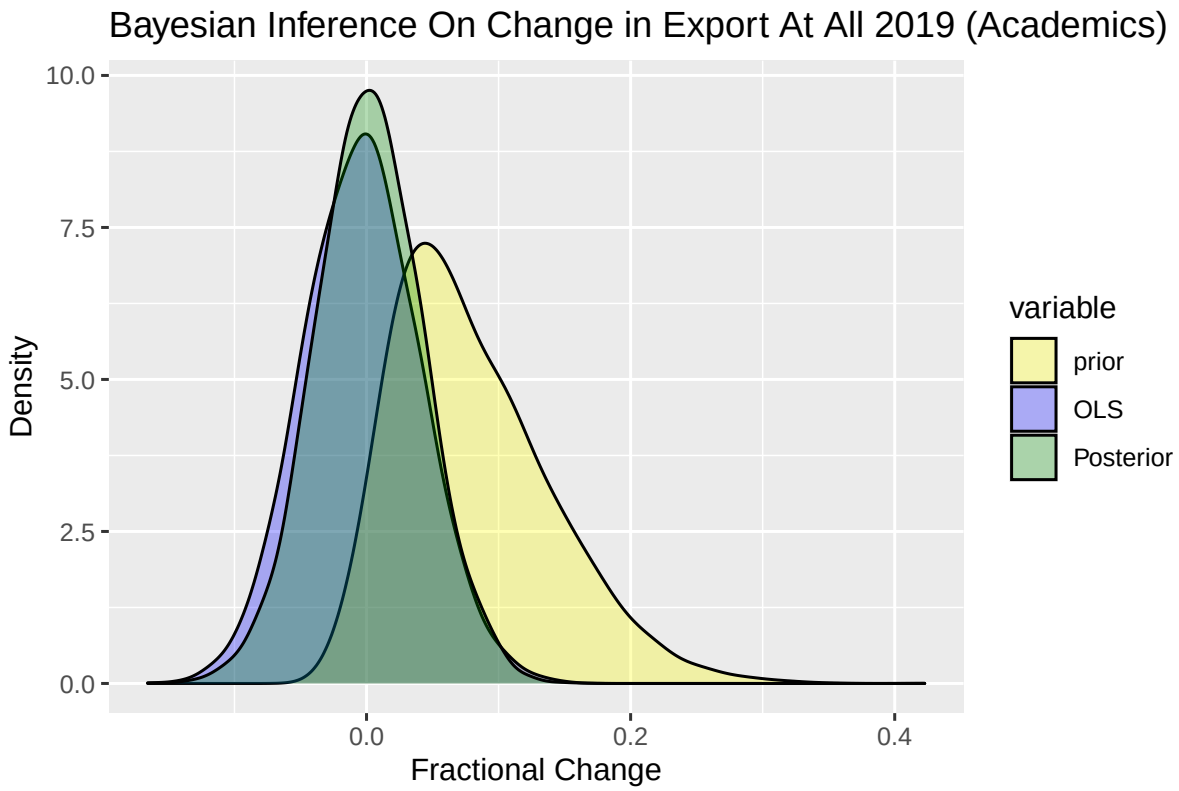
Panel B uses monthly data and clusters standard errors at the firm level.

Panels A and B use data on formal employment from the PILA, supplemented in column 3 of panel A by cross-checks from phone calls and the RUES.

Panels C through E use data from the RUES. I.H.S. denotes inverse hyperbolic sine transformation.

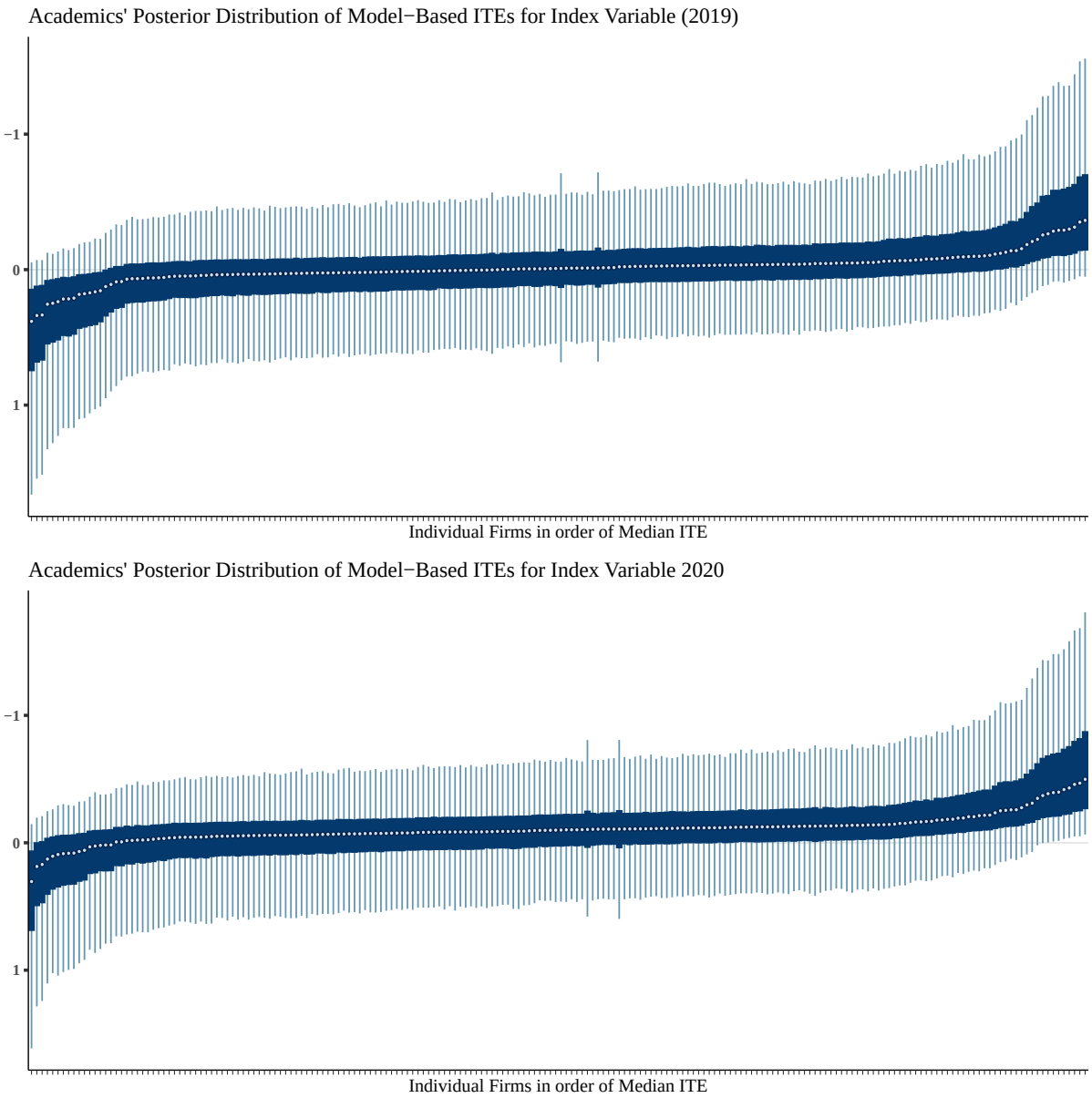
F Additional Bayesian Figures

Figure A8: Graphical Bayesian Updating of Academics' Priors



Notes: The prior displayed in this graph is the distribution fit to the elicited prior belief draws from the academics, as this is the input that actually goes into the posterior.

Figure A9: Bayesian Model-Based Distribution of ITEs Academics



Notes: For each ITE the dark blue band shows 50% credible interval and the line is the 95% credible interval. These inferences are based on the Gaussian potential outcome distribution using elicited priors.

G Unforeseen Challenges for Bayesian Inference in our Project

In this section we discuss several unforeseen complications with implementing Bayesian inference using our elicited priors on our data. The primary issue that we did not anticipate is that eliciting priors on the raw outcome scale (e.g. probability for binary variables, counts for count variables) would present substantial difficulties when fitting non-linear models such as logit for binary variables or non-negative binomial (or Poisson) for count data, because the coefficients are not on the raw scales and have nonlinear relationships to their raw-scale counterparts. To convert a prior distribution on a treatment effect expressed in probabilities (or counts) into a prior distribution on logit coefficients (or changes in rates) requires a nonlinear transformation of variables that moreover must make reference to a base probability or rate which is not perfectly known. Nonlinear transformation of variables is a challenge especially in the absence of autodifferentiation capability (as is the case in Stan) and introducing uncertainty onto the base probability is a substantial additional complication. Our initial efforts to simplify the problem by treating this as a transformation of parameters, picking a single base probability and hoping this might be "good enough", were not fruitful, leading to badly-behaved MCMC sampling and untrustworthy results.

The problem was even worse for variables we had assumed would be distributed continuously, because in fact there turned out to be large discrete spikes of data at zero in every case. Consider a simplified distribution consisting of a spike of zeroes added to a positive continuous tail (which is often called the "slab" because it looks flat compared to the large modal spike in the PDF). While in general one can handle this kind of data by using a "spike and slab" likelihood model (as discussed in various sources including Chapter 8 of Imbens and Rubin, and Meager 2022), in our case, this causes our prior on the average treatment effect to map to many possible priors on the parameters of the likelihood. The fundamental problem is that when there is a spike and slab structure to the data, there is an extensive and intensive margin of any change to the distribution. In our simplified example, a positive ATE in a spike and slab model may arise when an intervention "moves" firms in the spike into the positive tail, or when the mean of the positive tail is itself "moved up", or a combination of both types of effects. This means the mapping of the prior on the treatment effect in raw terms to priors on the coefficients in the spike and slab model is one-to-many and thus indetermined without additional information. We believe that it is possible to elicit such information by asking about the subjective probability that any average effect is produced by an extensive or intensive margin effect, but we did not foresee this data structure and did not ask this question.

We believe that future work will be able to make inroads into these important problems

now that they have been identified as a barrier to a more comprehensive Bayesian analysis of this kind of data. As our initial attempts to resolve the problems in an intuitive and computationally tractable manner were not fruitful, we were not able to solve this problem within the timeframe devoted to this project, and therefore we failed to complete some of the analyses we had laid out in our pre-analysis plan.

The spike and slab structure is also why we were not able to report quantile treatment effects for any of our data despite our intention to do so; conventional asymptotics for quantile inference require continuous underlying distributions, which we do not have. Binary and count data are also not continuously distributed and require substantial additional work to produce quantile inference (see for example Machado and Santos Silva 2005 on "jittering").