Submitted to *Econometrica*

BAYESIAN IMPACT EVALUATION WITH INFORMATIVE PRIORS: AN APPLICATION TO A COLOMBIAN MANAGEMENT AND EXPORT IMPROVEMENT PROGRAM

LEONARDO IACOVONE
The World Bank

DAVID MCKENZIE
Development Research Group, The World Bank

RACHAEL MEAGER
Department of Economics, University of New South Wales

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 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 the confidence intervals. Finally, for

Leonardo Iacovone: liacovone@worldbank.org
David McKenzie: dmckenzie@worldbank.org
Rachael Meager: r.meager@unsw.edu.au

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; two anonymous referees and our seminar audiences for helpful comments; Andrew Gelman for feedback on our pre-analysis plan; Robert Grant, Ed Jee and Brice Green for help with Stan; 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 Registry on June 26, 2018 (AEARCTR-0003109) and accepted as a stage 1 registered report at the Journal of Development Economics.

noisy outcomes like export value, posterior intervals show almost no updating from priors, highlighting how uninformative the data are about such outcomes. Future policy experiments could use these posteriors as priors in a Bayesian or empirical Bayesian analysis.

KEYWORDS: Bayesian Impact Evaluation, Prior Elicitation, Randomized Experiment, Management.

1. INTRODUCTION

Governments and researchers often test new policies by experimenting on a relatively small number of units. This is particularly common with expensive government programs intended to help small and medium enterprises. Typical studies in the literature work conduct multi-year experiments on 50-200 firms.[1] 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.

Yet 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 constraints facing beneficiaries. Academics bring knowledge from the existing literature and economic theory. 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 ignore.

Bayesian analysis theoretically offers a principled way of incorporating this prior knowledge into impact evaluation. Modern statistical texts such as Spiegelhalter et al. (1994), Gelman et al. (1995), Berry (2006) and Imbens and Rubin (2015) emphasize the utility of

[1] For example, Bloom et al. (2013) piloted a management improvement program in India with 28 plants; Higuchi et al. (2017) a program with treatment groups ranging from 26 to 133 firms in Vietnam; Bruhn et al. (2018) a program with 150 treated firms in Mexico; Iacovone et al. (2022) a program with a 159 firms in Colombia; and Custódio et al. (2020) a program with 93 firms in Mozambique.

a Bayesian approach to randomized trials. Efron (2012) emphasizes the natural use of empirical Bayesian methods when there are multiple experiments. Yet 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 demonstrates that using informative priors in Bayesian analyses is feasible and useful for policy experiments in practice. We consider an evaluation of a program in Colombia that was designed to help increase exports by improving management practices. This multi-million dollar experiment on a sample of 200 firms tackles a key policy question - how to diversify exports and increase firms' productivity. We elicit full prior distributions from academic experts, Colombian policymakers, and participating firms, and find that these priors are relatively informative. 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. We also consider priors we directly calculate from relevant pre-existing literature. We provide a standard frequentist analysis for comparison, as well as a textbook-style Bayesian analysis using diffuse priors, to show how much of the results are coming from the Bayesian methodology versus the information contained in the elicited priors. Our approach is quite general and could be applied to future experiments in this or other literatures, with a natural extension to the case of multiple experiments via empirical or hierarchical Bayes (Efron (2012)).

Our experimental results show the advantages of the informative Bayesian approach in a real-world policy setting. Frequentist inference in our setting finds small and statistically insignificant impacts of the program on exporting in 2019 (the year after the intervention started), and statistically significant negative impacts on some export outcomes in 2020. These estimates differ from the positive impacts of the program anticipated by academics, policymakers, 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 posterior intervals overlap almost completely with traditional confidence intervals. Even with informative priors, the signal in the data is strong enough to completely overturn the priors. 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. Frequentist estimates of the impacts of the program on the value and productivity of exports are very imprecise, with wide confidence intervals; the corresponding Bayesian posterior intervals show almost no updating from the priors, highlighting how uninformative the data are about such outcomes. Finally, for an overall index of export performance, our Bayesian posterior intervals lie between the prior intervals and confidence interval, showing partial updating. The Bayesian results show the program was not as successful as anticipated, but also temper the conclusion of it perversely worsening export outcomes.

The Bayesian framework also offers several natural extensions to the analysis above which are particularly relevant to policy. For example, we can calculate the posterior probability that the program had an impact large enough to pass a pre-specified threshold for scaling up. We find this probability is small ($< 10\%$) for most outcomes. We can also estimate the distribution of "individual treatment effects" (ITEs) by directly modelling the potential outcomes. We find little potential for heterogeneity.

This paper contributes to 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 (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 been elicited as means, at most with variances attached; they have not usually been elicited as full distributions, which allows us to capture potential skewness, bimodality, etc. Nor have they been used as priors for Bayesian 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.

Future experiments on similar policies could use our posteriors as their priors either in a Bayesian or empirical Bayesian framework, though, as noted by Efron (2012), "similar" is always a subjective consideration on some level, so context-specific elicited priors will continue to be useful even as a literature grows large. Bayesian meta-analysis approaches have already been used to aggregate evidence from multiple experiments either within the same study or across different studies (e.g. Meager (2019); Vivalt (2020)); taking the Bayesian approach to impact evaluation directly also offers the chance to update these posteriors in each additional experiment, and learn gradually as a literature evolves (Berry (2006)). The relative precision (or lack of it) in our results can also guide future experiments, either informally or formally within a Bayesian evidence-based decision-making framework (Abadie et al. (2023)). Finally, if prior elicitation becomes more common in applied economics, the properties of these priors can be studied in themselves, and the approach can be improved or refined.

2. CONTEXT, EXPERIMENT AND INTERVENTION

2.1. *The program*

The Colombia Productivity and Export Improvement Program (PEIP)[2] had the explicit objective of “improving the productivity and export capacity of the participating firms”. These two goals are important because Colombia’s exports are currently highly dependent on a small number of commodities such as crude petroleum, coal, coffee, flowers, and gold; and labor productivity is low. 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). The program aimed to increase exports through providing consulting services to firms to improve their management practices. The program was launched in November 2017, and applications closed March 23, 2018 (see timeline in Appendix A; all appendices are in the online supplement, Iacavone et al. (2025)). Both the treatment and control groups received a diagnostic from consultants valued at $625, and then the treatment group was offered 190 hours of consulting services valued at $13,800. Appendix B of the same document provides more details of the content and implementation of these consulting services.

[2] This is a pseudonym, which we use at the request of the Program of Productive Transformation (PTP)

2.2. *Random Assignment and Firm Characteristics*

510 firms applied for the program, of which 200 submitted all required documents and met the eligibility conditions. These 200 firms were then randomly assigned to two groups of 100. The application form data were used to stratify firms by size (small, medium, or large), and whether the firms had exported at all in the last 3 years. An additional two strata were added: one of export outlier firms, and one of a single firm that was missing firm size information. Within each of the eight strata, we ranked firms by an index of the proportion of 11 export practices they were using, and formed quadruplets, randomly selecting two firms from each quadruplet to be assigned to control, and two firms to treatment. This gives 54 strata defined by the export practice quadruplets inside the eight original strata.[3]. This process of forming many strata follows the standard experimental approach to improving balance and statistical power by trying to make the treatment and control groups as balanced as possible on key variables thought to be predictive of export performance.

Table I provides summary statistics for program firms. At the time of applying, 58% had exported in the last three years. 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. Firms had room to improve their export and general management practices. On average, firms were doing 36 percent of basic export practices, and 44 percent of general management practices. The firms are very heterogeneous, making it more difficult to detect treatment impacts. This heterogeneity is evident in firm size, with firms having a mean of 73 and median of 42 employees, but ranging from 2 to 750 workers. The annual sales of a firm at the 90th percentile ($8.6 million) are more than five times those of a firm at the median ($1.5 million), and 36 times those at the 10th percentile ($235,000).

3. ELICITING PRIORS: WHAT, HOW, AND FROM WHOM?

Using informative priors in Bayesian impact evaluation requires finding a way to bring in outside knowledge external to the data collected. This may come from any existing related academic literature or results of similar previous experiments (in which case one can use

[3] 3 of these 54 strata end up having a single firm in them each. These single firms then do not contribute to identification of treatment effects in our frequentist analysis which includes a dummy for each stratum.

Empirical Bayes). Often the existing literature may have few, if any, studies with a similar-enough intervention and context, as in our case. One can then elicit priors from key experts and decision-makers. This requires deciding from whom to elicit these priors, how to elicit them, and how to process the resulting information for input into the analysis.

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. Most experiments will face partial compliance, so researchers must decide whether to obtain priors on the intention-to-treat (ITT) effect or on the local average treatment effect (LATE) for the primary outcomes.

In some respects and contexts, the LATE may be a more natural object of interest, since it captures the impact of a program on those who actually take it up when offered, assuming the exclusion restriction holds. However, it is also more demanding to estimate and form beliefs about. The ITT is somewhat less demanding on both fronts, and typically still of interest for policy, especially if takeup is high.[4] This conceptual simplicity also makes the ITT potentially more comparable across individuals and experiments, as even if all the LATE conditions are satisfied in one setting or in the mind of some individuals we survey, this may not be the case in others. Thus, 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, Iacavone et al. (2025) for language used).

We also asked respondents about 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 high as anticipated. Second, one of our two Bayesian approaches uses a joint likelihood over the sets of potential outcomes with and without treatment. This inference depends on the inherently unobservable correlation between a firm's export

[4] If takeup is likely to be low, then this decision (ITT vs LATE) will require much more careful consideration.

performance with the program and without. We therefore also asked our academic experts for their priors on this correlation.

3.2. *How should priors be elicited?*

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 so as not to impose undue constraint on the shape of their beliefs during the elicitation process (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. We do this separately for each outcome, since eliciting the joint distribution of treatment effects across multiple outcomes would be extremely complex. 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 our Appendices Iacavone et al. (2025) for an example).[5]

A further question that might arise when eliciting these distributions for use in 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

[5] The raw expression of the priors therefore can have intervals with zero mass, which if taken literally could be problematic, but as we discuss in subsection 3.4 we will later fit a smooth distribution to the "stones" in order to use the output which will also take care of this issue.

responses.[6] Instead, researchers may do better by ensuring the language used in elicitation surveys 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. 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. It could be interesting for future research to examine the sensitivity of policymakers' priors to their knowledge of how these priors will be used.

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: (1) policymakers involved in designing and implementing the intervention. They should have the best knowledge of the local context and policy aims. Posteriors based on their priors should then be directly useful for making decisions; (2) domain-relevant academics who can combine their knowledge of the existing literature with how much they expect it to translate to the new context; (3) program participants, who apply to programs with priors about how much the program will help them.

We collected priors between June and October 2018, as the program was in the diagnostic phase. This timing was chosen so that the details of the intervention were as clear as possible, while being before policymakers and firms could see any program effects. Priors were collected from seven high-level Colombian policymakers, 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.

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

There are two reasons for collecting priors from multiple respondents – in our case, from 7 to 11 respondents per group. First, as we aim to use the priors to improve the estimation and inference, the wisdom of crowds suggests that the group average prior may be more accurate 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 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 interested in obtaining priors for distinct types of users (policymakers, academics, program respondents), since there are typically multiple decision-makers involved in using knowledge. In many practical settings there may only be a small number of policymakers involved in decisions, or global experts with deep knowledge of a topic. One can then examine how sensitive the results are to omitting the priors of any single expert, as illustrated in Appendix D, Figure D3 (Iacavone et al. (2025)).

One concern that might occur is whether individuals with a vested interest in an intervention may provide biased priors to try to game the system, or because they are over-optimistic.[7] Several points are worth noting. First, priors are transparent and the direction of their influence is often relatively obvious. Second, pre-registering priors makes them available for critique and discussion before they are used, independent of results. Third, there are many possible sources of priors including previous research which cannot be easily gamed. Berry (2006) and U.S. Department of Health and Human Services Food and Drug Administration (2010) documents how Bayesian approaches are used in regulatory decision-making in other fields, including both literature and expert elicited priors.

[7] A reviewer gave the example of whether the Food and Drug Administration (FDA) should consider priors provided by drug companies. This is an interesting area for applied decision-theoretic research, though the FDA already does consider such priors: U.S. Department of Health and Human Services Food and Drug Administration (2010). They encourage the use of literature priors; in the case of elicited expert priors they suggest contacting the FDA before the trial itself is done, which seems aligned with a notion of pre-registration.

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. But since we use Markov-Chain Monte-Carlo (MCMC) methods to simulate draws from unknown distributions for our Bayesian analysis, we require analytic PDFs.[8] 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. Pairwise Kolmogrov-Smirnov tests of equality of distributions typically reject the null of no difference in prior distributions between the academics, policymakers, and firms. There is considerable overlap in the distributions, but we typically see the academics to be more pessimistic about the likely impacts of the program than policymakers, with firms somewhere in between.[9] We will see that this pessimism leads the academics to have priors that are the closest to the data of these three groups in our setting, but whether academics would continue to be the most accurate in giving priors in other experiments with interventions which turn out to have large positive impacts is an open question.

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

[8] It is still valuable to avoid parametric structure at the elicitation stage, to avoid constraining the shape of their beliefs, and allow for skewness and multimodality. We then chose the analytic families for the PDFs after viewing the histograms of their elicited output.

[9] Table II reports the medians of the elicited priors for each outcome and expert group, and Figures 1-4 graphically show 95 percent intervals from these distributions.

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.[10] Based on the existing literature on business training programs and management consulting (McKenzie and Woodruff, 2014), 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. More details are provided in Appendix D (Iacavone et al. (2025)). In practice these literature-informed priors overlap substantially with our elicited priors, providing further credence in these elicited priors.

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 when 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-Student t, half-Normal or half-Cauchy priors on variance parameters (Gelman (2006)). 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.

[10] This is analogous to the idea of forming a surrogate index to predict long-term outcomes given short-term results (Athey et al., 2019), where here the literature may only provide prior information on an intermediate outcome, that we then want to extrapolate to the final outcome.

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 export products and destinations, and improve export performance. 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, 2022) 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. 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) (Ministry of Health and Social Protection, 2022), 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 (Iacavone et al. (2025)) defines each outcome in more detail, and replication data are provided in Iacovone et al. (2024).

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:

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

where $Y_{i,t-s}$ is the *s*th pre-intervention lag of the outcome of interest; δ_j are randomization strata fixed effects; 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 the treatment effect being equal to the medians of the prior distributions of policymakers, academics, and firms.

4.3. *Estimation of Treatment Effects using Bayesian Methods*

Our headline Bayesian analysis takes as its starting point the regression in equation 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. 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 tractability.[11] This produces the following likelihood model:

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

We now need to place priors on each of the parameters that govern the likelihood. We perform the analysis using five different priors on the key parameter of interest β, as follows: the elicited priors from academics, policymakers, and firms; the literature prior; and the

[11] This choice does not affect comparison with our frequentist estimates since homoskedastic standard errors are almost identical to our reported heteroskedastic-robust errors.

default highly diffuse prior. 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. The large number of strata fixed effects present a practical challenge in Bayesian analysis, as nuisance parameters are not partialed out as in OLS. As inference is done jointly on all parameters, a large number of nuisance parameters can result in overfitting and more uncertainty on β. We address this via regularization: on all δ_j, except a small set of key strata which we leave to vary freely, we place a hierarchical model $\delta_j \sim N(\delta, \sigma_\delta^2)$ in order to shrink them to their shared mean and then we regularize this mean towards zero using relatively tight priors on $(\delta, \sigma_\delta^2)$ with peak mass at zero. Further discussion can be found in Appendix F (Iacavone et al. (2025)) .

We also need priors on the other control variable coefficients and other parameters in the likelihood. We generally use weak, diffuse priors for these but 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 key 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.[12] These choices correspond to the "weakly informative" type of default prior approach seen in previous work, as discussed in section 3.6.

To calculate the joint posterior over all unknown parameters, we use a Hamiltonian Monte Carlo algorithm implemented via Stan in R. In each case we examine the performance via the effective sample size, traceplots and R-hat criterion.

We report the central 95% posterior interval on the treatment effect β as our headline result for each outcome. We can also compute other quantities of interest, such as the probability that β is large enough for the program to pass a policy threshold for scale-up; we also compute and report these results in section 5.3. In addition, once we have specified a

[12] Following Gelman (2006), which uncovered problems with the (inverse) Gamma priors on scale parameters, it is best to use Student t if there is sufficiently informative data, but for noisy estimation environments such as ours the Normal regularises more than the Student t and can be preferred, as in Gabry et al. (2019).

proper likelihood on the data, we can consider jointly modeling the potential outcomes to get more detailed causal inference on the distribution of treatment effects – we pursue this as an extension, and find little heterogeneity (Appendix G, Iacavone et al. (2025)).

The approach described above differs from the way that Bayesian statisticians or analysts in other fields would tend to approach this same data. They would typically uses richer likelihood models, 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-at-all is binary, etc). We use 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 priors.[13]

5. RESULTS

5.1. *Take-up, usage, and improvements in business practices*

Take-up and usage of the program was high. 88 of the 100 firms assigned to treatment started the intervention, and 83 have positive hours of consulting recorded. 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. We use our own firm survey to test whether this consulting improved export-oriented management practices (such as participating in trade fairs and getting quality certified for export markets), and general management practices (such as operations and commercial practices). Appendix B.3 (Iacavone et al. (2025)) shows no significant impact on export practices (with a point estimate of -4 p.p. relative to a a control mean of 59%), and a significant improvement of 6.7 p.p. in general management practices, relative to a control mean of 64 percent. The main improvements were in operations practices such as lean manufacturing methods, as well as commercial practices involving market research and customer interaction.

[13] There are also substantial challenges with implementing the analysis with the richer likelihoods, which we document in Appendix H (Iacavone et al. (2025)) .

5.2. *Impacts on export outcomes*

Figure 1 plots the trajectory of means for the different export outcomes by treatment status over the period 2010 to 2020. 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 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 II provides our frequentist estimates of the ITT effects on these different export outcomes, using the specification in equation 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 II. 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.).

The left panel of Figure 2 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 the central 95% interval of the different prior distributions in light blue, and then 95 percent posterior intervals from each corresponding posterior. First, note that the posterior 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 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 G, figure G.1 (Iacavone et al. (2025)) .

Next, panels B, C, D and E of Table II examine impacts of the program on the number of products exported, the number of countries the firm exports to, the number of unique product-country combinations exported, and whether the firm has exported a new product-country combination (export innovation). 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.

The right panel of Figure 2 provides the associated Bayesian impact evaluation results for export variety[14]. 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

[14] Figures **??** and **??** shows the results for the number of products and number of countries, and for export innovation.

the data shows a more negative impact, these stronger priors mean that the posterior only partially updates 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 use 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. We hoped bringing in prior information might result in narrower posterior intervals if these priors were consistent with the data.

Panels F and G of Table II 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 3 provides the Bayesian impact evaluation results. With a diffuse prior, the posterior intervals again are similar to the frequentist output. However, with informative priors, we see very little updating of the elicited priors. As the data here are not very informative, the posteriors place much more weight on the priors. 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 II 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.[15] Figure 4 shows the Bayesian results using the literature, policymaker and academic priors. We see substantial updating towards the data, with the posterior intervals centering around zero in 2019 and around small negative impacts in 2020.

5.3. *Bayesian Decision Analysis*

The Bayesian framework offers several options for additional analysis and inference directed towards policy decisions. For example, we can compute the probability that the program's average effects pass the minimum threshold necessary for policymakers to consider scaling it up. This threshold is known to the policymakers – possibly based on cost-benefit considerations,[16] but possibly also based on direct policy goals of export diversification and SME growth. During our prior elicitation exercise with the policymakers, 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.[17]

The results of this exercise for all types of priors are shown in table III 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; typically much less than 10% chance, and in some cases, virtually zero chance. The chances are typically lower in 2020 than in 2019, as expected given the movement of the data towards more negative outcomes. The exceptions are export value and productivity, where chances can be much higher - but this is primarily because we have little updating away from the priors even in 2020. The data is much more informative on the other five outcomes, all of which show little chance that the effect is large enough to merit scaling.

[15] 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.

[16] Although we note that this calculation would be complex to do formally, as it requires extrapolating treatment effects into the future and discounting them by some rate

[17] Because we have used Markov-Chain Monte Carlo methods to characterize the posterior distributions, and these methods simulate draws from these posteriors, the desired probability is just the proportion of draws from the posterior that lie above the threshold.

Second, as decision-makers may wish to go beyond the average effects, we can use Bayesian modelling of potential outcomes to infer the distribution of individual treatment effects. Appendix G (Iacavone et al. (2025)) discusses how we can do this by also eliciting priors on the correlation of outcomes in treatment and control states. We find at most a limited potential for heterogeneous effects.

5.4. *Why were impacts not as large as expected?*

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 examine different reasons for this lack of impact in our working paper (Iacovone et al., 2023), and 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. It is possible that some changes will take longer to manifest themselves, but at least in the first two years after the intervention, there is no sign of the program improving exporting.

6. CONCLUSIONS

This paper shows that Bayesian analysis using informative priors is feasible and useful for policy experiments. We highlight some of the practical issues involved in eliciting such priors, such as which treatment effect is to be estimated when there is incomplete compliance, who to elicit priors from, how to gather and fit these priors, and issues with large numbers of covariates or strata. 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 are overly influential. By comparing our estimates to those using a weak (default) prior, we also see that it is the informative priors and not other Bayesian modeling choices that drives results.

We see this approach as especially applicable for expensive long-term experiments with a limited sample size or when noisy outcomes limit statistical power. For example, in the case of export variety, where the informative priors are in line with the ex-post estimates, the resulting 95 percent posterior intervals are only 55-68 percent as wide as traditional confidence intervals. This can also be useful in larger trials with treatment effect heterogeneity, as many such studies are underpowered to detect interactions (Gelman (2018)). A similar problem arises in trials with multiple treatment arms. Informative priors on parameters, including the extent and dimensions of heterogeneity, could help overcome these power problems. The systematic learning and estimation advantages of the Bayesian or empirical Bayesian approach are particularly salient with multiple experiments, or as a given body of evidence grows over time (Efron (2012)). 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. [4]

ABADIE, A., A. AGARWAL, G. IMBENS, S. JIA, J. MCQUEEN, AND S. STEPANIANTS (2023): "Estimating the Value of Evidence-Based Decision Making," *arXiv preprint arXiv:2306.13681*. [5]

ANDREWS, I. AND J. SHAPIRO (2021): "A Model of Scientific Communication," *Econometrica*, 89, 2117–2142. [4]

ATHEY, S., R. CHETTY, G. IMBENS, AND H. KANG (2019): "The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely," *NBER Working Paper*. [12]

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. [8]

BERRY, D. A. (2006): "Bayesian clinical trials," *Nature reviews Drug discovery*, 5, 27–36. [2, 5, 10]

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. [2]

BLOOM, N., R. SADUN, AND J. V. REENEN (2016): "Management as a Technology," *Stanford Working Paper*. [5]

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. [2]

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*. [12]

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*. [2]

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. [8]

DELLAVIGNA, S. AND D. POPE (2018): "Predicting Experimental Results: Who Knows What?" *Journal of Political Economy*, 126, 2410–2456. [4, 9, 10]

DELLAVIGNA, S., D. POPE, AND E. VIVALT (2019): "Predict science to improve science," *Science*, 366, 428–429. [4]

DIAN (2022): "Unpublished data on Exports 2010-2020," *Dirección de Impuestos y Aduanas Nacionales*. [13]

EFRON, B. (2012): *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press. [3, 5, 22]

GABRY, J., D. SIMPSON, A. VEHTARI, M. BETANCOURT, AND A. GELMAN (2019): "Visualization in Bayesian Workflow," *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182, 389–402. [15]

GELMAN, A. (2006): "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)," . [12, 15]

——— (2018): "You need 16 times the sample size to estimate an interaction than to estimate a main effect," *Stat Modeling Blog*. [22]

GELMAN, A., J. B. CARLIN, H. S. STERN, AND D. B. RUBIN (1995): *Bayesian data analysis*, Chapman and Hall/CRC. [2]

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. [4]

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*. [2]

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. [4]

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. [9]

IACAVONE, L., D. MCKENZIE, AND R. MEAGER (2025): "Supplement to 'Bayesian Impact Evaluation with Informative Priors'," *Econometrica Supplemental Material*. [5, 7, 8, 10, 12, 13, 15, 16, 18, 21, 28]

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. [2]

IACOVONE, L., D. MCKENZIE, AND R. MEAGER (2023): "Bayesian Impact Evaluation with Informative Priors : An Application to a Colombian Management and Export Improvement Program," *World Bank Policy Research Working Paper*, 10274. [21]

——— (2024): "Replication package of "Bayesian Impact Evaluation with Informative Priors : An Application to a Colombian Management and Export Improvement Program"," *Zenodo*, http://doi.org/10.5281/zenodo.14430028. [13]

IMBENS, G. AND D. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press. [2]

LEMOINE, N. P. (2019): "Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses," *Oikos*, 128, 912–928. [12]

MANSKI, C. (2004): "Measuring expectations," *Econometrica*, 72, 1329–1376. [8]

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. [4]

McKenzie, D. and C. Woodruff (2014): "What are we learning from business training evaluations around the developing world?" *World Bank Research Observer*, 29, 48–82. [12]

——— (2017): "Business Practices in Small Firms in Developing Countries," *Management Science*, 63, 2967–2981. [11]

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. [5, 12]

Ministry of Health and Social Protection (2022): "Unpublished data from the PILA 2010-2020," *Planilla Integrada de Liquidaci/'on de Aportes a Seguridad Social*. [13]

Otis, N. (2022): "Policy Choice and the Wisdom of Crowds," *Working Paper*. [10]

Spiegelhalter, D. J., L. S. Freedman, and M. K. Parmar (1994): "Bayesian approaches to randomized trials," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 157, 357–387. [2]

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. [12]

U.S. Department of Health and Human Services Food and Drug Administration (2010): "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials," *Center for Devices and Radiological Health Division of Biostatistics Office of Surveillance and Biometrics*. [10]

Vivalt, E. (2020): "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economics Association*, 18, 3045–3089. [5]

TABLE I

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 II

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:*		
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 (I.H.S.)**		
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:			*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 (I.H.S.)**		
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:			*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:			*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)	0.000	0.000			

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.

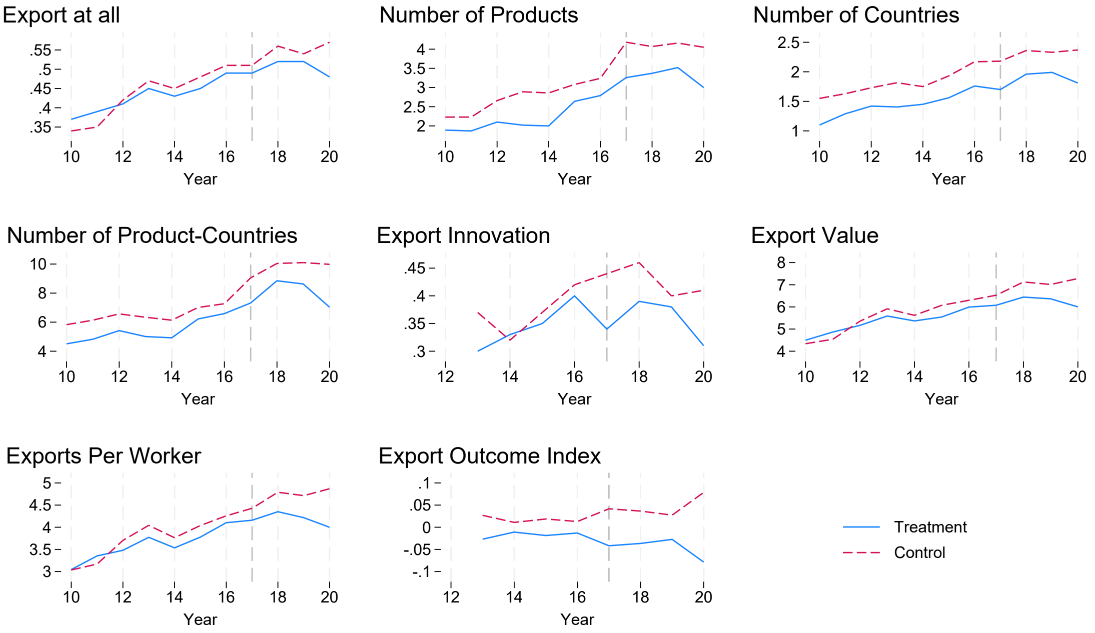
TABLE III

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.

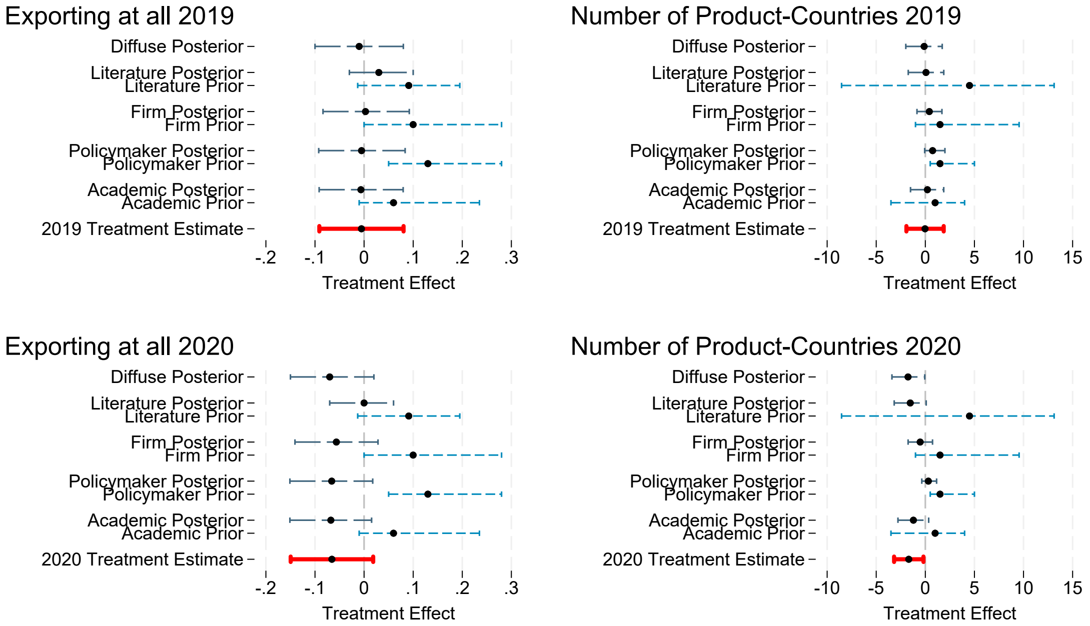
FIGURE 1.—Trajectory of Means of Primary Export Outcomes by Treatment Status



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 (Iacavone et al. (2025)) for variable definitions.

Co-editor [Name Surname; will be inserted later] handled this manuscript.

FIGURE 2.—Frequentist and Bayesian Estimation of Treatment Impacts on the Extensive Margin of Exporting at All and on Number of Varieties



Exporting at all 2019

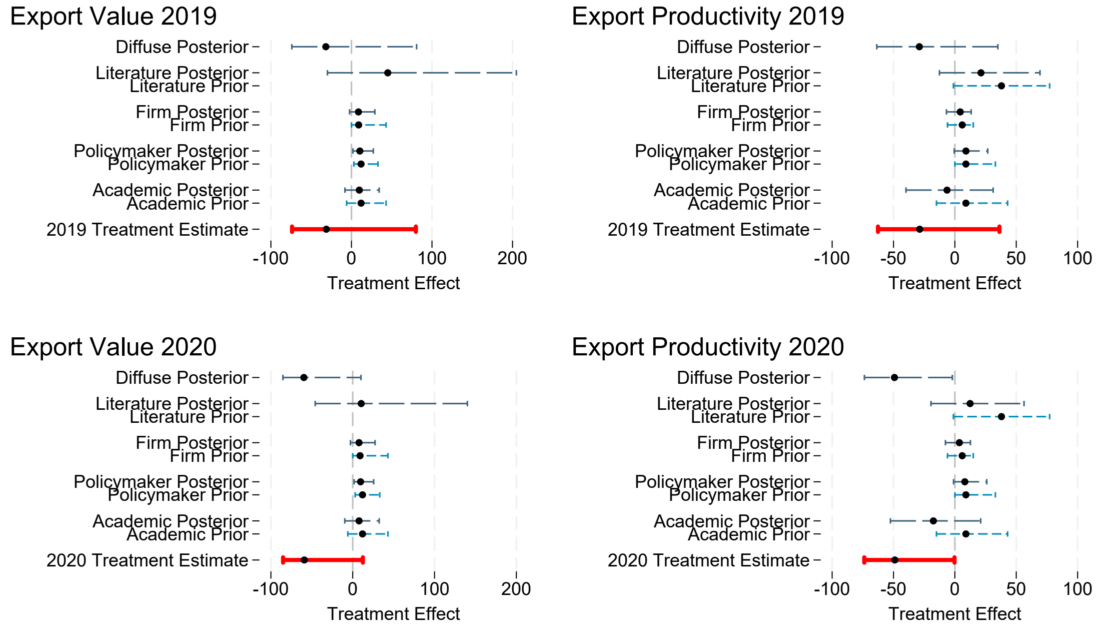
Number of Product-Countries 2019

Exporting at all 2020

Number of Product-Countries 2020

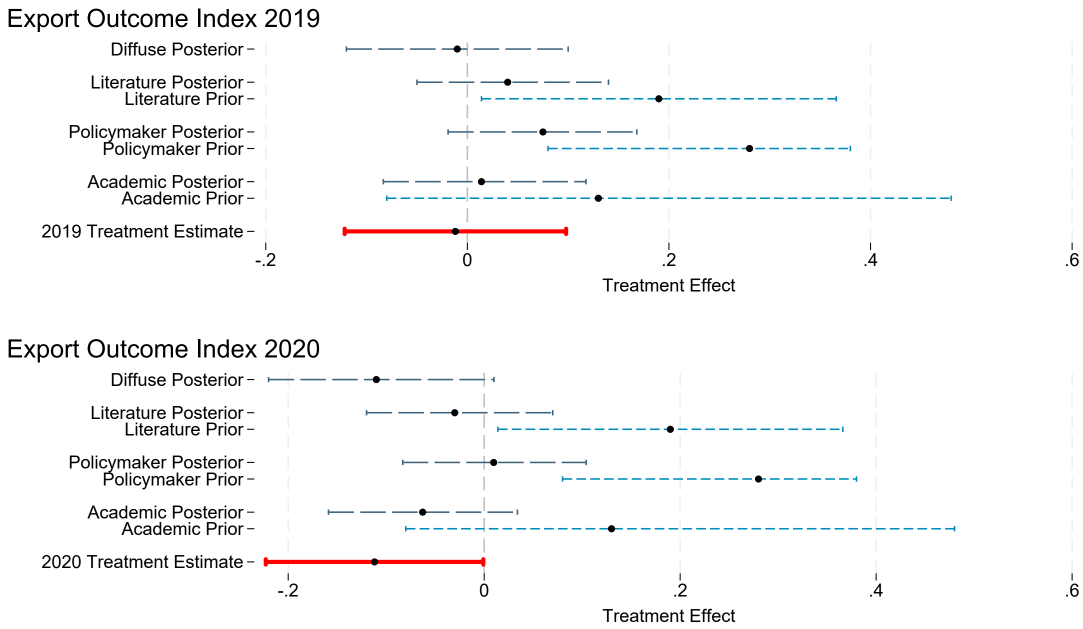
Notes: Treatment estimates show frequentist ITT estimate and associated 95 percent confidence intervals. Treatment effect units are in terms of the change in the proportion of firms exporting. Short dashed lines show 95 percent prior intervals elicited from academics, policymakers, and firms, and derived from the literature, with circles indicating median of prior distributions. Bayesian posteriors and long dashed 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. Control mean for exporting at all is 0.54 in 2019 and 0.57 in 2020. Control mean for product-country varieties is 10.1 in 2019 and 10.0 in 2020.

FIGURE 3.—Frequentist and Bayesian Estimation of Treatment Impacts on Export Value and Export Productivity



Notes: Treatment estimates 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. Short dashed 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 long dashed 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 4.—Frequentist and Bayesian Estimation of Treatment Impacts on Overall Export Performance Index



Notes: Treatment estimates 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. Short dashed 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 long dashed 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.