SUPPLEMENT TO
"Bayesian Impact Evaluation with Informative Priors: An Application to a Colombian Management and Export Improvement Program"
by L. Iacavone, D. McKenzie and R. Meager

APPENDIX 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 both treatment and control
The program started implementation in June 2018 and finished in December 2019.
(Most firms received the intervention between July 2018 and July 2019.)
Timing of Data Collection
June 2018-October 2018: Prior Elicitation from Policymakers, Academics, and firms.
November 2019-May 2020: Follow-up survey of Management Practices.

APPENDIX B: MORE DETAILS OF THE SAMPLE AND CONSULTING INTERVENTIONS

B.1. *Firm Characteristics*

Table I provides means of baseline variables by treatment status, along with percentiles of the baseline distribution to illustrate some of the heterogeneity. 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; and a company making dried tropical fruits.

The key assumption underlying equation **??** is the stable unit treatment value assumption. 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, 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, most of the more common product categories have additional within-category heterogeneity (e.g. cotton t-shirts, miscellaneous plastic products). Our assumption is therefore that any export growth from the treated firms is unlikely to be primarily business stealing from control firms.

B.2. *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. 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).

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. Our working paper (Iacovone et al., 2023) provides more details of the consulting intervention.

Table B.I 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, which 79 firms 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.

TABLE B.I

TAKE-UP RATES BY AREA

Area	Number of Firms Using	Hours Mean	conditional SD	on 25th	using 50th	Area 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

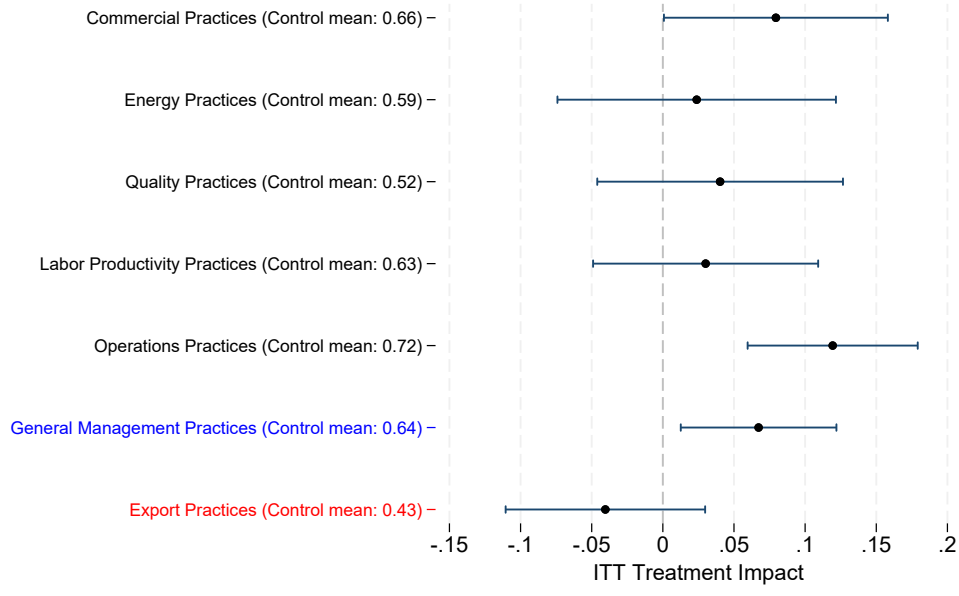
B.3. *Impacts on business and export practices*

The most immediate impacts of management consulting interventions are typically seen on management practices, as firms implement the changes recommended by consultants. There

are 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. 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). Our main measures of export practices and management practices are then defined as the proportion of these practices that are used by firms.

Figure B.1 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 **??**. 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.

FIGURE B.1.—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).

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 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 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.

APPENDIX C: DATA AND MEASUREMENT DETAILS

We use an exchange rate of U.S. $1 = 2985 Colombian pesos for reporting quantities in dollars in the text.

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 and are outlined in more detail in Iacovone et al. (2023).

APPENDIX D: MORE DETAILS ON PRIOR ELICITATION AND LITERATURE PRIORS

We provide further details on the priors here, with full details contained in our pre-registration in the AEA registry (https://www.socialscienceregistry.org/trials/3109).

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 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

Number and width of bins

Several practical considerations arise when eliciting prior distributions of treatment effects 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). Delavande et al. (2011) shows elicited distributions appear to be relatively robust to the number of stones and bins used, and to whether bin ranges are self-anchored or pre-determined. They find highest precision for using pre-determined bins and 20 stones, and this is what we use. Our bins fully cover the range of prior beliefs, with the top and bottom bins receiving no mass for 6 out of our 8 outcomes. Only the top bin for export labor productivity (<0.5% of mass for academics), and the top bin for export innovation (10.7% of mass for policymakers) had non-zero use.

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]

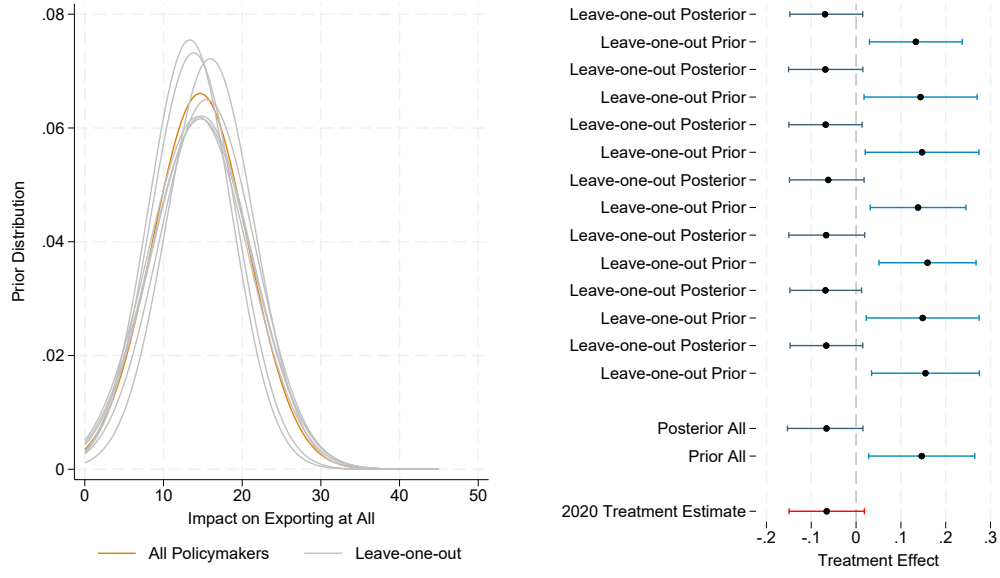
BELIEFS ABOUT THE IMPACT ON THE PERCENT OF FIRMS THAT EXPORT

-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

Robustness to any single expert In many cases the number of experts with deep knowledge of the program, or policymakers with decision-making power can be relatively small. This may raise concerns about how sensitive results are to the priors of a particular decision-maker. Figure D3 illustrates how one can examine this. The left panel shows that the fitted prior from using all 7 policymakers is similar to those obtained by any combination of 6 policymakers and dropping one. The right panel shows that as a result both the prior intervals and posterior credible intervals are very robust to leaving out any single decision-maker.

Obtaining Literature Priors McKenzie and Woodruff (2017) show that an approximate estimate of the treatment effect of business and management training on firm outcomes is equal to the treatment effect on an index of business practices, multiplied by the cross-sectional correlation between business practices and a given firm outcome. We construct a baseline index of export practices as the proportion of 11 different measures of whether the firm is doing activities such as participating in trade fairs, planning production in time for export markets, getting quality certifications, and segmenting clients by international market. McKenzie and Woodruff (2014) report that business training programs typically increase the proportion of business practices implemented by 0.05 to 0.10. We take 0.10 given the intervention is more intensive than most business trainings. This also lines up with the 9 percentage point increase

FIGURE D3.—Examining Robustness to Dropping Any Single Policymaker in Forming Priors



in management practices found in Colombia by Iacovone et al. (2022). We assume a standard deviation of the impact on export practices of 0.03 to reflect the typical variability seen across these studies. We then regress baseline outcomes on our export practices index, and use these associations in forming the literature priors. For example, to get the literature-informed prior on the extensive margin effect of exporting at all, we multiply the association between exporting at all and export practices in 2017 (0.907) by our prior on the treatment effect on export practices (0.10) to get a mean for the literature prior of 0.091. We obtain a standard deviation for this prior we assume independence of the errors in the effect on export practices from the error in the association between export practices and export outcomes to calculate a standard deviation of 0.05 in this case. The AEA registry provides full details.

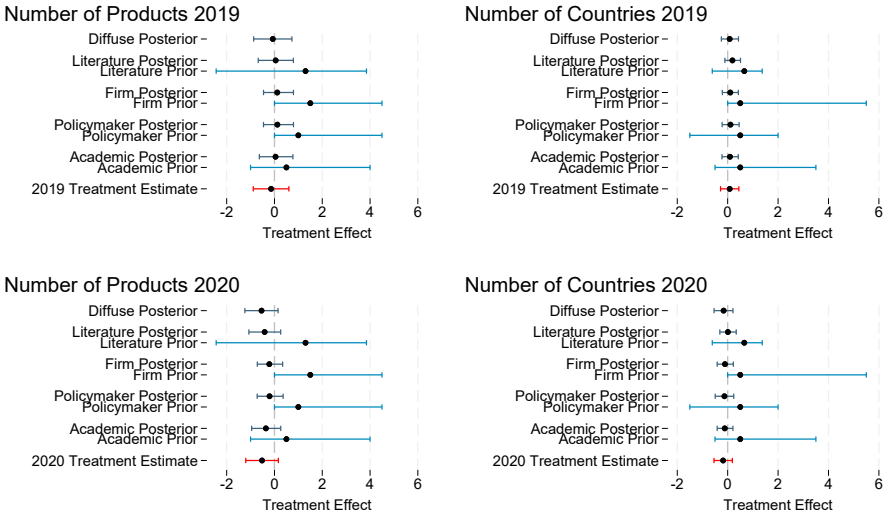
APPENDIX E: IMPACTS ON OTHER OUTCOMES

Figures E1 and E2 show the frequentist and Bayesian impacts on number of products, number of countries, and export innovation (exporting a new product-country combination). Our working paper also examines impacts on secondary outcomes of interest such as sales, employment, survival, and productivity (Iacovone et al., 2023). We did not elicit priors on these secondary outcomes, and so do not conduct Bayesian analysis of these impacts. There is a marginally significant impact of 3-5 percentage points on survival, which appears to come from smaller firms that do not export. We do not find any significant impacts on other outcomes.

APPENDIX F: DEALING WITH MANY STRATA DUMMIES IN BAYESIAN ESTIMATION

In translating the frequentist approach into a Bayesian model, one issue which arises is the large number of 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

FIGURE E1.—Impacts on Number of Products and Countries Exported

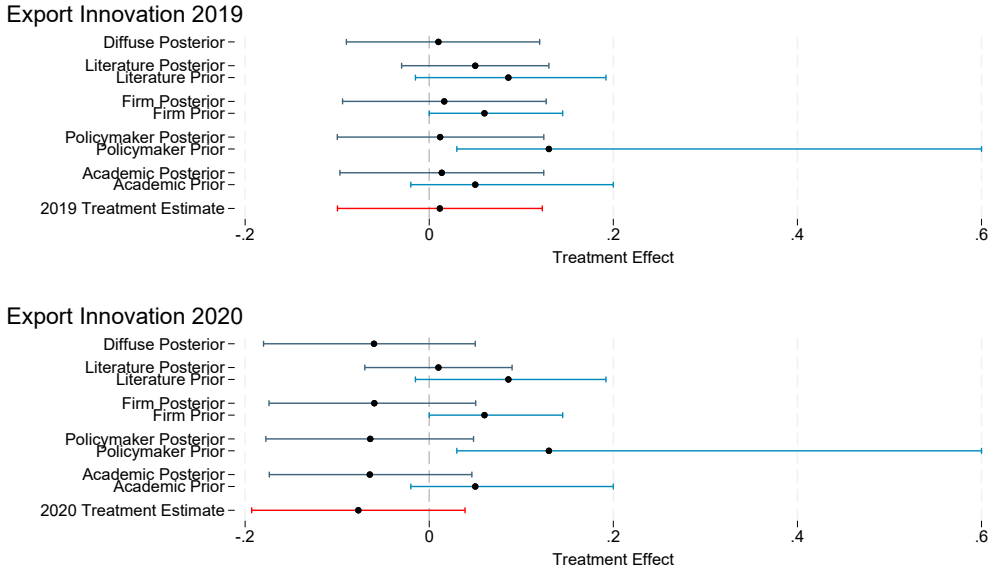


Notes: Treatment estimates and associated red line show frequentist ITT estimate and associated 95 percent confidence intervals. Treatment effect units are in terms of the change in the number of products or countries. Control Mean is 4.16 products and 2.33 countries in 2019, and 4.05 products and 2.37 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..

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

FIGURE E2.—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.

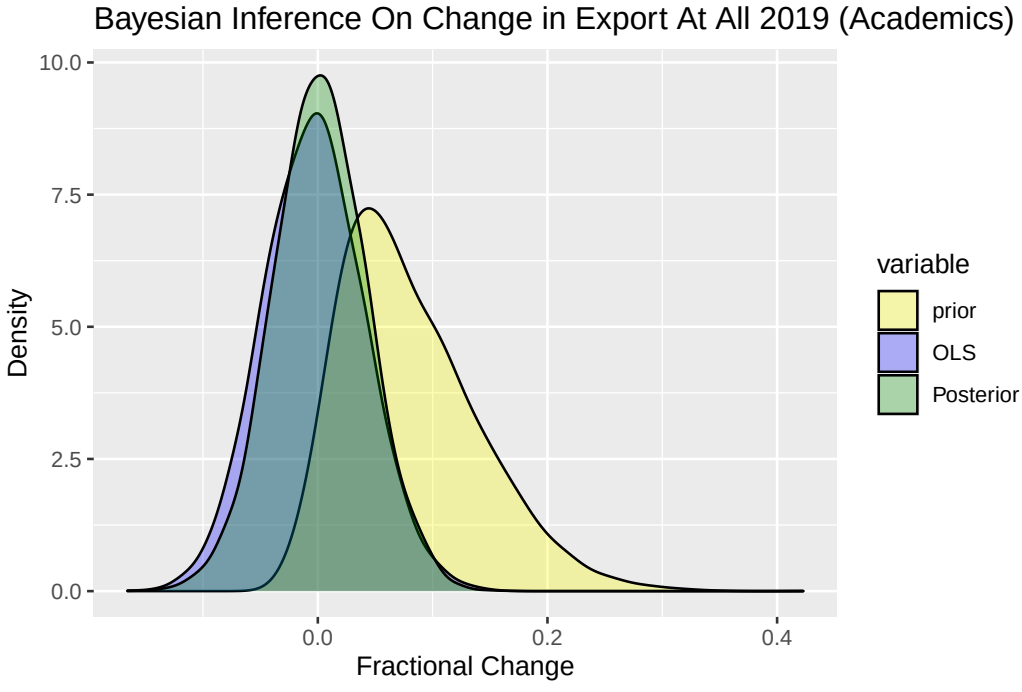
with the treatment assignment variable, they will be able to overcome the penalty imposed by the hierarchical structure and the priors. Note however that we do not constrain the role of the 3 main factors that drive the stratification: firm size (implemented via 3 categories), exports in last 3 years 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.

APPENDIX G: ITEs AND ADDITIONAL BAYESIAN FIGURES

We can also use our Bayesian framework to estimate the conditional distribution of individual treatment effects. Such inference requires a joint likelihood $f(Y(0)_i, Y(1)_i|X_i)$ of some kind. 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 with mean $X_i\beta$ in generic econometric notation.[1]. 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 choice is the bivariate Gaussian. We 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 (as in Imbens and Rubin (2015)). For brevity we will now denote the conditional expectation of Y_i given only

[1]The two outcome models' expected values differ only in switching on or off the binary treatment indicator covariate (as discussed regarding equation **??**)

FIGURE G.1.—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.

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:

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

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.[2]

Any inference - frequentist or Bayesian - on the distribution of individual treatment effects (ITEs) in the sample requires some information about ρ. The data never contains such information, since by definition we never observe both $Y(0)_i$ and $Y(1)_i$ for any i, so it is natural

[2]Heterogeneous effects can arise even without explicitly modelling heterogeneity in β. Even when β is just a number, because of correlations in the unmodelled variation in the potential outcomes; that is, when ρ is not zero.

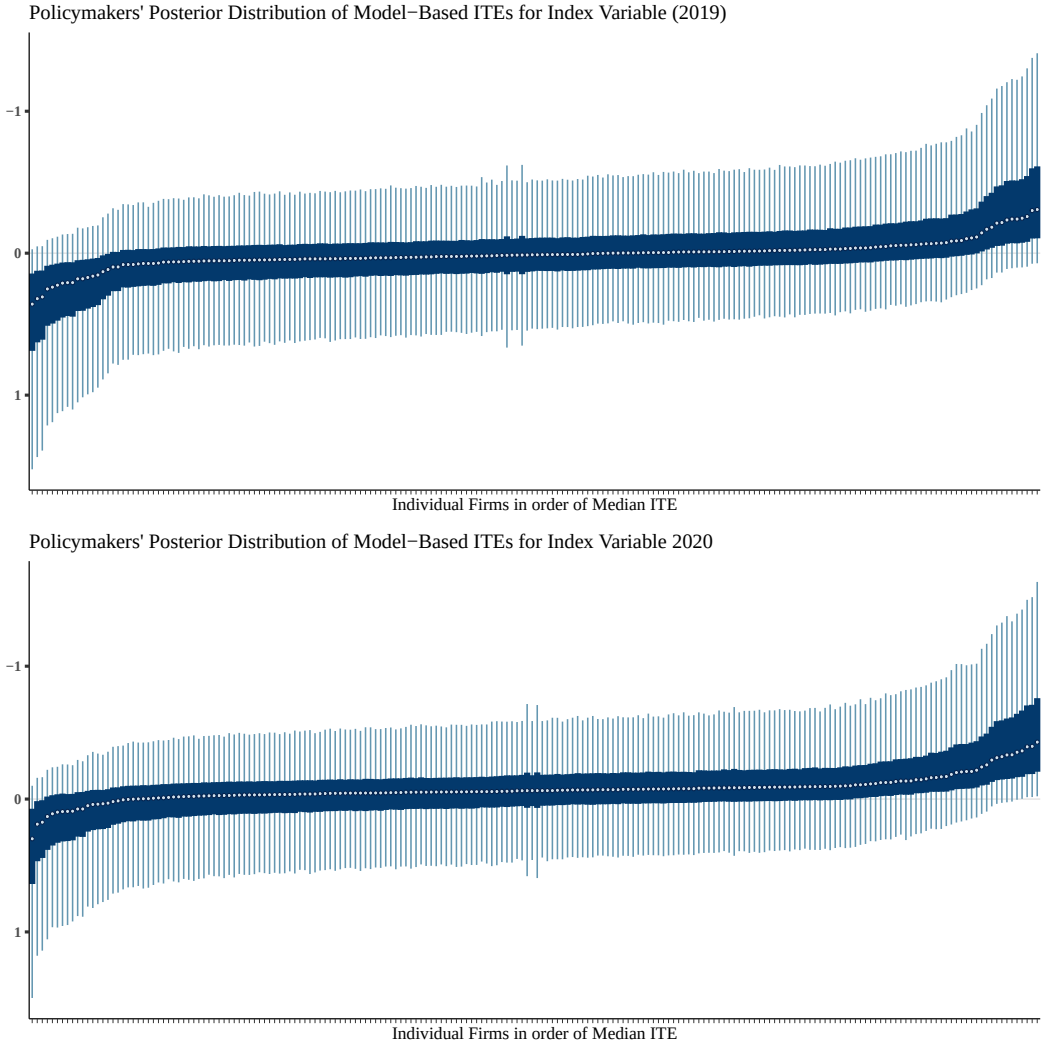
in the Bayesian set-up to use priors. We elicited beliefs about ρ from academics and policymakers using the same methods as used when we elicited priors about β. Both academics and policymakers believed that firms doing well under the control regime (the status quo) would also do well under treatment, implying ρ is likely positive. Combining this with the data and the model in equation 1 allows us to infer the distribution of individual treatment effects (ITEs) with appropriate posterior uncertainty.

The results for the Index outcome data are shown in figure G.2 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 some chance that certain 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. But the majority of the firms in the sample were unlikely to have seen any results from the intervention. This is a natural 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, as shown in Iacovone et al. (2023).

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 symmetry in the Gaussian likelihood. The real outcome data is not symmetric – it has a discrete spike at zero and a positive continuous tail. More complex likelihood models could accommodate this, but in our case the priors on treatment effects will have a one-to-many mapping to priors on parameters of such a model.[3] This is a general issue for our Bayesian analysis preventing the use of more complex likelihoods without substantial additional effort; further discussed in appendix H. Yet we still find the analysis informative, as the results overall suggest only a limited potential for heterogeneous effects.

[3]This would arise for example in a spike-and-slab 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.

FIGURE G.2.—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.

APPENDIX H: UNFORESEEN CHALLENGES FOR BAYESIAN INFERENCE IN OUR PROJECT: DEVIATIONS FROM OUR PRE-REGISTRATION

In this section we discuss several unforeseen complications with implementing Bayesian inference using our elicited priors on our data, which explain why we have not carried out all the inference we wished to, according to our Pre-Registration document. 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 (the "slab"). 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 (2015), and Meager (2022)), in our case, this causes our prior on the average ITT 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 average effect 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.

The spike and slab structure is also why we were not able to report quantile treatment effects for any of our data despite our pre-specified intention to do so; conventional asymptotics for quantile inference require continuous underlying distributions, which we do not have. Likewise, binary and count data also require substantial additional work to produce quantile inference (see for example Machado and Santos Silva 2005 on "jittering"). We believe that future work will be able to address these issues now that they have been identified as a barrier to a more comprehensive Bayesian analysis in practice.

APPENDIX: REFERENCES

ATHEY, S. AND G. W. IMBENS (2017): "The econometrics of randomized experiments," in *Handbook of economic field experiments*, Elsevier, vol. 1, 73–140. [8]

DELAVANDE, A., X. GINE, AND D. MCKENZIE (2011): "Eliciting Probabilistic Expectations with Visual Aids in Developing Countries: How sensitive are answers to variations in elicitation design?" *Journal of Applied Econometrics*, 26, 479–497. [6]

GELMAN, A., J. B. CARLIN, H. S. STERN, AND D. B. RUBIN (1995): *Bayesian data analysis*, Chapman and Hall/CRC. [8]

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. [8]

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. [7]

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. [2, 4, 7, 11]

IMBENS, G. AND D. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press. [9, 10, 13]

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. [6]

——— (2017): "Business Practices in Small Firms in Developing Countries," *Management Science*, 63, 2967–2981. [6]

MEAGER, R. (2022): "Aggregating distributional treatment effects: A Bayesian hierarchical analysis of the microcredit literature," *American Economic Review*, 112, 1818–1847. [13]