Cover page

Journal of Development Economics

Registered Report Stage 1: Proposal

What are the Effects of Improving Management Practices on Exporting Among Colombian SMEs? A Comparison of Bayesian and Frequentist Impact Evaluation Approaches

Leonardo Iacovone, David McKenzie, Rachael Meager and Darío Rodríguez Pérez

Date of latest draft: 06/26/2018

Keywords: Management, Exports, Bayesian Impact Evaluation, Randomized Experiment

JEL codes: O12, F14, C11, C93

Study pre-registration: Please provide the registry name, registration identifier, and/or a link to the public registration if available. If a pre-registration has not been completed, please indicate whether you plan to complete pre-registration before the commencement of data collection. See the 'Pre-registration Resources' section of the JDE RR Author Guidelines for a list of registries.

Proposed timeline *(required)*

If granted accepted based on pre-results review, please indicate when this study will be completed?

June 2020.

Abstract *(required)*

Many developing country governments seek to improve the productivity and export competitiveness of their SME sectors. We will run a randomized experiment to test whether improving management practices can achieve these goals, in the context of the Colombia Productiva program. A treatment group of 100 firms will receive a diagnostic of their management practices, 190 hours of technical assistance, and then participate in a trade fair, while a control group of 100 firms will receive the diagnostic and trade fair only. Rich administrative data on export transactions will enable us to track whether this program leads to firms being more likely to export, diversifying what they export and

where they export to, and improving export productivity. A key methodological innovation is the comparison of a Bayesian impact evaluation framework to a frequentist approach, highlighting what we can learn from each in a sample of modest size.

Reporting checklist for Stage 1 submissions *(optional)*

Reporting checklist table with research design details available for download here.

1. Introduction

Research question: background, importance and relevance

- What is the main problem/question motivating the study? Why is this question important for the field of development economics?
- How has this problem/question been addressed thus far in the relevant literature? What are the competing theories for explanation of this question? How is this study different from prior research on this problem/question?

This project is motivated by both a substantive policy question and a methodological research question. A key policy priority in many developing countries is to diversify and expand the export base. This is the case in Colombia, which is currently highly dependent on commodities such as crude petroleum, coal, coffee, flowers, and gold, which make up more than 80 percent of its merchandise exports. The *Colombia Productiva* program aims to broaden the range of firms and sectors that engage in exporting, and aims to do so through improvements in the management practices through high-intensity technical assistance to small and medium enterprises (SMEs). The main research question is then whether such a program can improve export outcomes in Colombian SMEs?

We will answer this question using a randomized experiment, in which 100 SMEs will be assigned to receive a diagnostic, followed by 190 hours of technical assistance delivered over 10 months, and then the firms will participate in an export fair. This treatment group will be compared to a control group of 100 SMEs that will receive the diagnostic and participate in the export fair, but will not receive technical assistance. The firms are heterogeneous in size, ranging from 2 to 750 workers, with a mean of 73 workers. This combination of a modest size number of firms, of very heterogeneous types, is typical for SME development programs in many developing countries (McKenzie, 2011). However, it presents a challenge for traditional impact evaluation approaches that rely on null-hypothesis significance testing, since the statistical power for detecting a difference in the means of the two groups will be low. Power can be improved by stratifying on baseline data on the outcomes and eliminating strata with large outliers (Bruhn and McKenzie, 2009), as well as by incorporating multiple rounds of pre-intervention data to increase the time dimension of the sample (McKenzie, 2012). But even after these methods are used, the minimal detectable effect sizes for some outcomes may be larger than effect sizes of research and policy interest.

This paper therefore seeks to develop and implement a Bayesian impact evaluation approach as an alternative method of learning from policy experiments in which there is pre-existing knowledge or priors. We propose obtaining prior distributions for the impact of treatment from policymaker in the implementing agency, academics, participating firms, and the pre-existing literature. We will then use this prior distribution, and the evidence from the experiment, to calculate a posterior distribution. This can be used to calculate posterior means and other parameters of the posterior distribution to provide information on the probability that the program has had certain impacts, as well as how much the evidence from the policy experiment has shifted priors.

This research will make contributions to three literatures. The first is a literature on the importance of management practices for explaining differences in the performance of firms across and within countries (Bloom and van Reenen, 2007, Bloom et al. 2014). In contrast to a much larger number of experiments of business training for microenterprises (reviewed in McKenzie and Woodruff, 2014), there have only been a small number of evaluations of interventions to improve management in SMEs and larger firms. Bloom et al. (2013) find five months of intensive consulting improved management and led to productivity improvements in plants with an average of 130 workers in India, and Bruhn et al. (2018) find that a year of once a week consulting on firms with an average of 14 workers in Mexico improved productivity and, over a longer term, employment. Higuchi et al. (2015) find a mix of on-site and classroom Kaizen training in Vietnam helped firms with an average of 20 workers survive, and have higher value-added. Each of these studies also provides the intervention for a relatively small number of firms[1], and none of them study interventions explicitly tailored to exporting, nor look at exporting as an outcome. Nevertheless, Bloom et al. (2017) use data on American and Chinese firms, and find that better managed firms are more likely to export, sell more products to more destinations, and have higher export revenues, with their structural estimates suggesting that poor management practices may especially matter for export performance in less-developed countries.

The second literature concerns the impact of firm-specific policy interventions to increase export performance. To date there have been relatively few rigorous evaluations of these policies in developing countries. Several studies use ex-post evaluations with non-experimental methods (e.g. Girma et al. (2009) on production subsidies for Chinese firms, and Cadot et al. (2015) on matching grants given to Tunisian firms to implement export business plans), but concerns remain about self-selection of firms into these programs. There have been few experiments on interventions to spur exporting. A notable exception is Atkin et al. (2017), who carried out a demand-side intervention with small firms (average size of one employee), providing Egyptian rug-manufacturers with initial orders and links to foreign buyers, and find exporting increases firm productivity. Two other experiments do not find statistically significant impacts of much lighter information interventions: Breinlich et al. (2017) consider the impact of sending brochures from the export promotion agency to SMEs in the United Kingdom, and Kim et al. (2016) who find that one-day informational seminars had no impact on exporting for textile firms in Vietnam with an average of 35 workers. This existing literature has not examined whether interventions to improve management increase export performance.

Finally, this work will be the first example we know of using Bayesian impact evaluation for a prospective field experiment. Several recent studies have elicited expectations of treatment effects from policymakers, or experts, in order to assess the accuracy of these predictions (e.g. Groh et al, 2016; Hirschleifer et al.

[1] 12 plants were treated in Bloom et al, 2013; 150 firms were offered treatment in Bruhn et al. (2018), of which 80 took it up; and Higuchi et al. (2016) had multiple treatment groups, with 10-76 firms per group getting on-site assistance.

2016; Dellavigna and Pope, 2017), but do not use their priors for subsequent analysis. Vivalt (2017) conducts a Bayesian meta-analysis, eliciting beliefs from uninformed policymakers about the impact of conditional cash transfers on educational attendance (for studies that had already taken place), and then asking how much benefit each study would add when updating these priors. But, no our knowledge, there have been no examples of using Bayesian impact evaluation for evaluating economic policies. We use two approaches in designing and implementing such an approach. The first is to implement Bayesian specifications of our frequentist regressions. The second is to implement the Bayesian approach to model-based inference set out theoretically in Imbens and Rubin (2015).[2] Given a prior distribution of the treatment parameter, a model for the distribution of outcomes, and the observed data, one can derive the conditional distribution of the treatment effect estimand given the randomized assignment and subsequent observed data. This distribution can then be used to obtain the posterior mean, standard deviation, and quantiles of the treatment effect.

2. Research Design

Intervention(s)

- What type of an intervention does the study involve[3]? Elaborate in detail when, where and by whom it will be delivered. Please provide sufficient detail to allow for replication in line with this journal's Mandatory Replication Policy.
- How will individual observations be assigned to treatment and control conditions[4]?
- Are there multiple treatment arms involved and if so, are they exclusive or overlapping?
- What is the source of exogenous variation in your study?
- If applicable, what observations will be blinded (masked)[5] after assignment to interventions and how? If blinding is not possible, what measures will be taken to minimize the potential for performance and expectancy biases (e.g. keeping participants unaware of trial hypotheses, measuring participant and provider expectations of benefit at baseline, etc.)?
- The instructions and supporting materials for the administration of the intervention should be included as an appendix.

[2] Imbens and Rubin (2015) illustrate this approach by applying it retrospectively to the national supported work program, but use diffuse priors in their illustration since no priors were collected.

[3] For more information on reporting standards for interventions, see Hoffmann et al. (2014).

[4] For more information on what to report on randomization, see Bruhn and McKenzie (2009).

[5] Blinding or masking refers to methods of withholding information about assigned interventions post-randomization from those involved in the trial, when knowledge of this information could influence their behavior in a way that would later prove integral to interpreting the results (Grant 2017, 12)

The intervention consists of three distinct stages. The first, given to all firms, is a *diagnostic* stage. All firms receive a visit from a consultant who provides a diagnostic analysis of the business, and recommendations for improvements 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 second stage, given only to the treatment group, is the *intervention* stage. Here firms receive up to 190 hours of technical assistance: 30 hours directed towards general commercial strategy, and 160 hours towards the two of the other four areas in which they have the more room to improve. The diagnostic and intervention will be performed by consultants from a consortium of five Colombian consulting firms who specialize in the different areas. The final phase is a trade fair (*macrorrueda de negocios*) organized by *ProColombia* designed as a business matchmaking forum to enable firms to obtain meetings with international buyers, primarily from countries with which Colombia has free trade agreements. This will take place in the first half of 2019,

The treatment is estimated to have a market value of 40 million Colombian Pesos (approximately US$13,800). Small firms selected for the program have to pay 3 million COP ($1,035) and medium and large firms 6 million COP ($2,070), which can be paid in multiple installments.

To be eligible for the program, firms had to have existed for at least 2 years, be formally registered, belong to one of fourteen selected sectors (transport manufacturing, construction, textiles, fruits and fruit products, speciality coffees and coffee products, beef, aquaculture, cocoa products, processed food, cosmetics, pharmaceuticals, plastics and paint products, basic chemicals, and business process outsourcing/software), provide financial statements for the 2015-16 year and other documentation, and complete an online application process. The closing date for applications was March 23, 2018. A total of 200 firms applied for the program and met the eligibility criteria.

These 200 firms were then randomly assigned to two groups of 100 each on April 11, 2018. The application form data were used to stratify firms by size (small, medium, or large)[6], and whether or not the firms had exported at all in the last 3 years. An additional two strata were added: one stratum of 19 export outlier firms (defined in terms of having export values, the number of destinations exported to, or the number of different products exported above the 95th percentile in the self-reported export data on the application form), and one stratum of 1 firm that was missing firm size information. We then formed an index of the proportion of 11 exporting management practices (defined in Appendix A) that firms were using. Within each of the eight strata, we then ranked firms by this export practices index, and formed quadruplets, with two firms from each quadruplet assigned to control (benefits 1) and two firms to treatment (benefits 2). In total this gives us 54 strata defined by these export practice quadruplets inside the eight original strata.

[6] These categories are defined in Colombia by business assets. Small firms had between US$125,000 and US$1.25 million in assets; medium firms between US$1.25 million and US$7.5 million in assets; and large firms more than US$7.5 million in assets.

The program website, and the description given to firms made clear that the program had two benefit schemes (benefits 1 and benefits 2), and that firms would be randomly allocated by the World Bank to one or another using a process that guaranteed transparency and equality of opportunities for selection. Random assignment was carried out by two of the authors using Stata, and was livestreamed to both members of the Programa de Transformación Productiva, as well as to applicant firms. Firms are therefore not blinded to their treatment status, or to the existence of another treatment. This has the advantage that firms do not see their selection or not into the program as any signal of their management capabilities, since it was made clear that selection was done randomly. We seek to minimize the potential for bias by not describing the intervention as an experiment or test (and not using the labels treatment and control group when discussing the program publicly), keeping participant firms blind to our trial hypotheses, and relying largely on administrative data such as objective export performance data for our key outcome measures.

Hypotheses

- What are the main outcomes of interest? Which outcomes are primary to the analysis, which are secondary, and why?
- How will the main outcomes of interest be defined in your dataset?
- Please include all hypotheses which will be tested, linking each outcome specifically to how it will be measured. These should be reported as main results in the Stage 2 submission.

Primary Hypothesis: The Colombia Productiva program will lead more firms to export, diversity the range of products exported and destinations exported to, and improve the export performance of participating firms.

This primary hypothesis will be measured through the following primary outcomes, all obtained from administrative data on export performance, where *past year* will denote the year from the start of the intervention phase onwards (anticipated to be the period August 2018-July 2019):

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 one year period since the intervention begins, 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 will be coded as zero for firms that do not export, and will be 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 ($\log(y+(y^2+1)^{1/2}$) of total exports (measured in US dollars), and is 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. *A standardized export outcomes index:* This index will be calculated as the average of the normalized z-scores of outcomes 1 through 7, where each z-score is defined by subtracting the mean and dividing by the standard deviation of the respective outcome.

The use of the inverse hyperbolic sine transformation has two purposes. The first is that we believe it more plausible that the intervention will have a similar treatment impact on export values in percentage terms across firms of different sizes than a similar absolute impact. This transform is similar to a logarithmic transform, so that impacts can be interpreted in percentage terms. Secondly, the transformation improves power by reducing the influence of outliers.

Secondary Hypotheses

Secondary hypotheses examine the channels through which the intervention is expected to have an impact, as well as measuring impacts on additional outcomes of key economic and policy interest.

Secondary Hypothesis 1 (SH1): The Colombia Productiva program will improve the export-specific management practices used in firms, as well as their general business practices.

This will be measured through the following two secondary outcomes:

1. *Proportion of Export-Specific Management Practices Being Used:* An index that will range from 0 to 1, measuring the proportion of specific export practices being used by the firm. Appendix A shows the eleven such practices measured at baseline. We will refine this list by talking with the intervening agency and firms to better understand the key practices they see as essential to export performance, and register an updated definition in the AEA RCT registry before collecting this. This will be collected via a survey of the firms in the three months after the end of the intervention (anticipated to be June-August 2019). This will be coded as zero for firms that are closed.
2. *Proportion of General Management Practices Being Used:* An index ranging from 0 to 1, measuring the proportion of general management practices being used in the firm. Our baseline data contains 37 such practices, encompassing operations management, strategy, quality control, energy efficiency, productivity, and human resources practices. We will refine this list by using the protocols of the intervening agency to determine which types of practices they see as weak in the diagnostic phase, and where they hope to make changes. We will then register an updated definition in the

AEA RCT registry before collecting this. This will be collected via a survey of the firms in the three months after the end of the intervention (anticipated to be June-August 2019). This will be coded as zero for firms that are closed.

Secondary Hypothesis 2 (SH2): The Colombia Productiva program will improve the export competitiveness of firms by making them more energy efficient, lowering production costs, improving quality, and increasing productivity.

We aim to test this hypothesis by linking the firms to the annual manufacturing survey (EAM), and making an arrangement with the national statistics agency (DANE) to ensure that the firms in our study are included in this survey. The following secondary outcomes will then be measured for this hypothesis:

1. *Energy efficiency:* this will be measured as the ratio of energy costs to total production costs, measuring the amount of energy needed to produce one dollar of output. It will be winsorized at the 1st and 99th percentiles, and will be coded as zero for firms not operating.
2. *Production costs index:* this will be measured as a standardized index of z-scores of the following components: i) input prices index (the EAM measures prices of the key inputs, we will measure whether input prices drop); ii) total production costs as a ratio of sales (winsorized at the 1st and 99th percentiles); and iii) physical input usage index (the EAM measures quantities of the key inputs, and quantities of output – this will measure whether the firm is using fewer input units to produce a given unit of output).
3. *Labor productivity:* the inverse hyperbolic sine of sales per worker, coded as zero for firms not operating.
4. *Total factor productivity:* TFP, using input and output prices, and the method of Ackerberg et al. (2015) to measure productivity.

Three key points to note about these measures are that i) the EAM is collected annually, but released with delay. Data for the 2019 calendar year is therefore likely to only become available around October 2020; ii) we will at best have data for 2017, 2018 and 2019, so will have at most one year of pre-intervention data to condition on; and iii) these data can only be used in the data lab of DANE, and so will not be able to be publicly shared afterwards. Statistical power is likely to be lower for these outcomes as well. These outcomes are useful to look at, but given these limitations, are secondary to our main analysis.

Secondary Hypothesis 3 (SH3): The Colombia Productiva program will ultimately increase the employment, sales, and profitability of participating firms.

This will be measured through the following three outcomes.

1. *Employment*: monthly (formal) employment in the firm, winsorized at the 99th percentile, and coded as zero for firms not in business. This will be measured by linking firms to the PILA database, which collects administrative data on firms from their labor filings on workers.
2. *Inverse hyperbolic sine of annual revenue:* This will be obtained from the EAM, and expressed in Colombian pesos. It is coded as zero for firms not operating.

3. *Inverse hyperbolic sine of annual profits:* This will be obtained from the EAM, and expressed in Colombian pesos. It is coded as zero for firms not operating.

As with SH2, the revenue and profits outcomes will be subject to the same limitations on availability. The employment outcome will be more easily obtained from the administrative records.

Basic methodological framework / Identification strategy

- What is the basic methodological framework of the study (RCT, pre-post, simple comparison, difference-in-difference etc.) and why is it suitable to address this research question?

The study is conducted as a randomized experiment. We discuss in the methodology section below the frequentist and Bayesian approaches to estimation that are planned.

Data

Please use this section to provide details on pilot data and *prospective* data that you will collect after pre-results acceptance of your research design.

Sample

- What is the unit of analysis for this sample (individuals, organizations, etc.)?
- What is the expected sample size? If applicable, please include statistical power calculations[7] to justify sample size.
- What is the effect size you will be able to detect?

The sample consists of the 200 firms that applied for the program. Table 1 provides summary statistics for the firms by treatment group, using data from the application forms submitted by firms. The firms have been in existence for a median of 18 years, with 58 percent of them having exported at all in the last three years. On average, firms are doing 36 percent of the basic export practices measured on the application form, and 44 percent of general management practices. Figure 1 shows the distribution of these practices is similar for the treatment and control groups, and that the export practices distribution is right-skewed, with the mass of firms using very few practices. This suggests scope for most firms to improve.

The firms are very heterogeneous. This is evident in firm size (47 percent small, 45 percent medium, 9 percent large), with firms having a mean of 69 and median of 41 employees, but ranging from 2 to 750 workers (Figure 2). The sales of a firm at the 90th percentile (25,675 million pesos or US$8.6 million)[8] are more than five times those of a firm at the median (4,595 million pesos or US$1.5 million), and 36 times those at the 10th percentile (700 million pesos or US$235,000). This heterogeneity also shows up across

[7] Useful information and software tools for power calculations can be found here.

[8] 2970 Colombian pesos equals 1 USD.

sector, with the most common sectors being textiles (18%), construction (16%), transportation equipment (14%), plastics and paint products (13%), and processed food (10%). Almost half of the firms are located in the Cundimarca region that includes the capital city, Bogota, with another quarter in the two regions around Cali and Medellin.

After the random assignment, we used the official firm identifiers to match the firms in our study to official export record data. This provides us with data on the number of products (at the 6-digit level) exported, countries exported to, and export amounts for each year from 2010 through 2017. Figure 3 shows that the baseline (2017) distributions of our key export outcomes are also highly right-skewed. In 2017, exactly half of the firms had exported. Conditional on exporting, the median firm exports $170,000, exporting 3 different products, to 2 countries, and a total of 5 distinct product-country combinations. However, the means are much larger. Excluding the export outlier strata greatly lowers these means in the application form data (Table 1), but has less of an impact on the administrative export data (Table 2). We can also use the time-series nature of the data to trace the trajectory of exporting over time (Figure 4). This set of firms has been increasing exporting over this period, with 39 percent of the sample exporting to a new product-country combination in 2017 (that they hadn't exported to in the previous 3 years).

Using this pre-intervention export outcome data and the baseline export practices index, we run the following regressions to examine the association between export practices and export outcomes in our sample:

$$Exportoutcome_i = \alpha + \beta ExportPractices_i + \delta'^{X_i} + \varepsilon_i \qquad (1)$$

These regressions follow the same spirit as Bloom et al. (2017) and McKenzie and Woodruff (2017), and are intended for two purposes. The first is to check whether there is an association between export performance and better export management practices in our sample. Panel A of Table 3 shows that there is a positive and significant association between better practices and higher export performance for all export measures we consider: the extensive margin of whether firms export at all; the number of products, countries, and product-country combinations; the amount exported and export labor productivity (exports per worker); as well as measures of breaking into new product and country markets. Panel B shows that these associations continue to hold and are of similar magnitude when conditioning on general business practices (which are not predictive of better export performance after controlling for export practices). Panel C shows these associations remain, and are of similar magnitude still after controlling also for sector, firm age, firm size, number of employees, and region. The results therefore suggest a robust association between better export management practices and export performance.

The second purpose of this exercise is as an input into power calculations and into our research-informed priors. McKenzie and Woodruff (2017) suggest that the treatment effects in the existing business training literature are of the approximate magnitude obtained by multiplying the treatment effect on business practices from such trainings by the association between business practices and the outcome of interest. They also note that many trainings change the proportion of business practices being implemented by 0.05 to 0.10, with larger impacts observed in some of the more intensive training programs and consulting engagements. We therefore consider the likely magnitude we can expect to see if the Colombia Productiva program improves export practices by 0.10, which we obtain by multiplying the coefficients in Table 3 by 0.10. For example, this suggests we might expect to see the likelihood of a firm exporting at all increase by 0.10*0.859 = 0.086. This is the assumed intent-to-treat (ITT) effect. McKenzie and Woodruff (2014) note that the typical training program has 65 percent attendance, so the treatment effect on the treated is typically 1.5 times the ITT. We anticipate compliance of at least 80 percent in our program, so are being slightly conservative in assuming an ITT of 0.10 on export practices.

Table 4 then provides our power calculations. This is a context in which the baseline data from the application form, coupled with the eight years of pre-intervention data on exporting provides us with more data than is usually the case when calculating power ex-ante. The first rows of the Table provide the key parameters needed for power calculations. This includes the mean and standard deviation of export measures in 2017 (and export practices at the time of application in 2018), and the autocorrelations using our previous years of data. In addition, in order to examine the power gains obtained from stratification, we also report the residual standard deviation that remains after conditioning on the randomization strata. Our power calculations are for an intent-to-treat effect.

Consider, column 1, which examines our power to detect an improvement in export business practices (SH1). Given a baseline mean of 36 percent of practices being implemented, and our sample sizes of 100 treatment and 100 control firms, we have 89.5% power to detect a 10 percentage point improvement in practices using a simple comparison of means. The minimal detectable effect size is 9 percentage points, and this drops to 2 percentage points after controlling for stratification, and to 1 percentage point when also controlling for the baseline value in an Ancova specification (McKenzie, 2012). The result is that we are confident in our ability to measure whether the Colombia Productiva program improves the business practices of firms. This is in line with the existing literature, which have typically been able to detect even relatively small improvements in practices.

The remaining columns then examine our power to detect improvements in our primary hypothesis export outcomes (columns 2 to 7). In addition, column 8 considers log employment, to illustrate power for a key secondary outcome of policy interest (SH3). In contrast to export practices, we see that power to detect treatment effects of the magnitude suggested by the associations in Table 3 is extremely low if we just compare treatment and control means. For example, power to detect a 1.1 increase in the number of products exported (relative to the baseline mean of 3.7), is only 15.2 percent. Conditioning on the randomization strata leads to some improvements in power, but power is still below 80 percent in all cases. The main power improvements then come from being able to control for one or more rounds of pre-intervention data in an Ancova specification. The high autocorrelation of export outcomes is the key to power gains here. The one outcome considered in this table for which power will still be low is whether or not the firm exported a new product-country combination in the past year. This export innovation is not as strongly autocorrelated as export levels, making it harder to detect a treatment impact.

The bottom of Table 4 then shows the minimal detectable treatment effects to achieve 80% power, using different estimation methods. While a simple comparison of treatment and control means has very large

MDEs, using one or more rounds of the pre-intervention administrative time series data yields MDEs of a size that are likely to be of policy interest for most outcomes. The main exception is export value, where the MDE is still 49% (0.40 log points) after controlling for five rounds of data. This reflects the large dispersion in this outcome, as seen in Figure 3.

Table 1: Balance Table based on Application Form Data

	Full Sample		No-outlier Sample	
	(1)	(2)	(1)	(2)
	Benefits1	Benefits2	Benefits1	Benefits2
	Mean/SE	Mean/SE	Mean/SE	Mean/SE
Variables used in Stratification				
Small Firm	0.460	0.470	0.506	0.511
	[0.050]	[0.050]	[0.053]	[0.053]
Medium Firm	0.430	0.460	0.438	0.433
	[0.050]	[0.050]	[0.053]	[0.053]
Large Firm	0.100	0.070	0.056	0.056
	[0.030]	[0.026]	[0.025]	[0.024]
Outlier in export data	0.090	0.100	0.000	0.000
	[0.029]	[0.030]	[0.000]	[0.000]
Exported in the last three years	0.580	0.580	0.539	0.533
	[0.050]	[0.050]	[0.053]	[0.053]
Export Practices Index	0.362	0.376	0.348	0.355
	[0.022]	[0.022]	[0.022]	[0.022]
Other Variables				
Firm located in Antioquia region	0.170	0.140	0.169	0.144
	[0.038]	[0.035]	[0.040]	[0.037]
Firm located in Cundinamarca region	0.520	0.440	0.528	0.433
	[0.050]	[0.050]	[0.053]	[0.053]
Firm located in Valle de Cauca region	0.070	0.150	0.067	0.156
	[0.026]	[0.036]	[0.027]	[0.038]
Firm age (years)	20.051	20.440	18.472	19.300
	[1.394]	[1.392]	[1.260]	[1.372]
Firm sales in 2016 (millions of pesos)	11416	11853	9104	10329
	[2028]	[2660]	[1808]	[2744]
Firm profits in 2016 (millions of pesos)	399	659	369	593
	[77]	[200]	[71]	[214]
Export value in 2016 (1000s of USD)	4548	24420	253	157
	[2972]	[20629]	[70]	[59]
Number of destinations exported to	2.556	2.190	1.742	1.300
	[0.452]	[0.403]	[0.287]	[0.204]
Exports to more than 2 destinations	0.293	0.270	0.236	0.211
	[0.046]	[0.045]	[0.045]	[0.043]
Number of products exported	35.110	91.740	8.303	4.444
	[19.284]	[60.655]	[2.298]	[0.921]
Exports more than 5 products	0.280	0.270	0.236	0.211
	[0.045]	[0.045]	[0.045]	[0.043]
Number of employees in 2016	69.969	68.061	61.885	57.270
	[10.114]	[8.937]	[10.248]	[7.675]
Business Practices Index	0.456	0.432	0.444	0.420
	[0.017]	[0.014]	[0.017]	[0.014]
Sample Size	100	100	89	90

Note: no outlier sample excludes strata comprising baseline export outliers

Table 2: Baseline Administrative Data on Exports

	Full Sample		No-outlier sample	
	Benefits1	Benefits2	Benefits1	Benefits2
	Mean/SE	Mean/SE	Mean/SE	Mean/SE
Exported at all in 2017	0.510	0.490	0.472	0.444
	[0.050]	[0.050]	[0.053]	[0.053]
Number of products (6 digit) exported in 2017	4.180	3.390	3.101	1.911
	[0.800]	[0.955]	[0.676]	[0.614]
Number of products (2 digit) exported in 2017	1.750	1.480	1.472	0.878
	[0.292]	[0.343]	[0.250]	[0.199]
Number of countries exported to in 2017	2.180	1.720	1.719	1.144
	[0.345]	[0.308]	[0.296]	[0.199]
Number of product-country pairs exported in 2017	9.040	9.840	6.584	3.578
	[1.917]	[4.151]	[1.681]	[1.526]
Number of other firms exporting same product-country	0.710	0.640	0.607	0.478
	[0.125]	[0.118]	[0.124]	[0.107]
Free on board export value 2017 (1000s of USD)	336	341	235	148
	[81]	[145]	[67]	[54]
Exported a new product in 2017	0.360	0.250	0.326	0.211
	[0.048]	[0.044]	[0.050]	[0.043]
Exported to a new country in 2017	0.260	0.250	0.247	0.233
	[0.044]	[0.044]	[0.046]	[0.045]
Exported to a new country-product in 2017	0.440	0.340	0.393	0.289
	[0.050]	[0.048]	[0.052]	[0.048]
Sample Size	100	100	89	90

Table 3: Association between 2017 Export Outcomes and Export Practices Index

	Export at all	# Products exported	# Countries exported to	# Product-Countries	Export Value	Export productivity	Exported a new product	Exported to new country	Exported a new product-country
Panel A: No other controls									
Export Practices Index	0.859***	11.090***	6.726***	35.171***	12.815***	3.015***	0.669***	0.579***	0.776***
	(0.127)	(3.304)	(1.065)	(9.744)	(1.627)	(0.455)	(0.138)	(0.129)	(0.139)
Panel B: Controlling for General Business Practices									
Export Practices Index	0.839***	10.928***	5.738***	34.985***	12.605***	3.609***	0.661***	0.628***	0.821***
	(0.154)	(3.834)	(1.020)	(11.087)	(1.934)	(0.569)	(0.157)	(0.152)	(0.157)
Business Practices Index	0.055	0.446	2.721	0.512	0.579	-1.655*	0.023	-0.136	-0.125
	(0.236)	(4.849)	(1.838)	(14.014)	(3.006)	(0.875)	(0.232)	(0.228)	(0.234)
Panel C: Controlling for General Business Practices, Firm Size, Employees, Sector, Region and Firm Age.									
Export Practices Index	0.907***	13.063***	6.599***	44.765***	13.683***	3.830***	0.733***	0.653***	0.858***
	(0.141)	(3.994)	(1.109)	(13.934)	(1.777)	(0.588)	(0.150)	(0.156)	(0.154)
Sample Size	200	200	200	200	200	190	200	200	200
Mean	0.500	3.785	1.950	9.440	6.302	1.076	0.305	0.255	0.390

Notes: robust standard errors in parentheses. *, **, and *** indicate significance at the 10, 5, and 1 percent levels.
Firms that do not export are included as having value zero for all outcomes above. Export value is the inverse hyperbolic sine transformation of thousands of US dollar exports.

Table 4: Power Calculations

	Export				# Product-	New	I.H.S.	I.H.S.	
	Practices	Export at	# Products	# Countries	Country	Product-	Export	Export	Log
	Index	all	Exported	Exported	Pairs	Country	Value	Productivity	Employees
Parameters (from baseline data)									
Baseline Mean	0.36	0.490	3.72	1.94	8.21	0.39	6.53	1.08	3.65
Baseline S.D.	0.22	0.500	8.43	3.22	21.3	0.49	6.58	1.53	1.21
Residual S.D.	0.04	0.297	6.19	2.17	16.1	0.34	3.64	0.85	0.79
1-year Autocorrelation	0.80	0.82	0.91	0.94	0.96	0.54	0.88	0.88	0.95
Average 5-year autocorrelation	n.a.	0.71	0.85	0.87	0.87	0.51	0.78	n.a.	n.a.
Assumed Treatment Effect	0.10	0.085	1.1	0.67	3.52	0.078	1.3	0.30	0.12
as a percentage	*27.8*	*17.3*	*29.6*	*34.5*	*42.9*	*20.0*	*366.9*	*35.0*	*12.7*
Power:									
Comparison of means	0.895	0.184	0.152	0.312	0.215	0.203	0.287	0.283	0.108
After controlling for strata	1.000	0.525	0.242	0.588	0.340	0.368	0.714	0.704	0.189
Ancova with 1 year before	1.000	0.942	0.858	1.000	1.000	0.487	1.000	1.000	0.931
Ancova with 5 years before	1.000	1.000	0.999	1.000	1.000	0.619	1.000	n.a.	n.a.
MDE at 80% power									
Comparison of means	0.09	0.21	3.4	1.3	8.5	0.20	2.7	0.61	0.48
After controlling for strata	0.02	0.12	2.5	0.9	6.4	0.14	1.5	0.34	0.32
Ancova with 1 year before	0.01	0.07	1.1	0.3	1.8	0.12	0.7	0.16	0.10
Ancova with 5 years before	n.a.	0.05	0.6	0.2	n.a.	0.10	0.4	n.a.	n.a.

Notes:
n.a. denotes not available.
MDE denotes minimal detectable effect size
Residual S.D. is standard deviation after controlling for strata fixed effects
Assumed treatment effect is 0.10 effect size on export practices multipled by associations in Table 3.
One-year autocorrelation in export practices based on data from another ongoing management experiment in Colombia.

Figure 1: Distribution of Management Practices

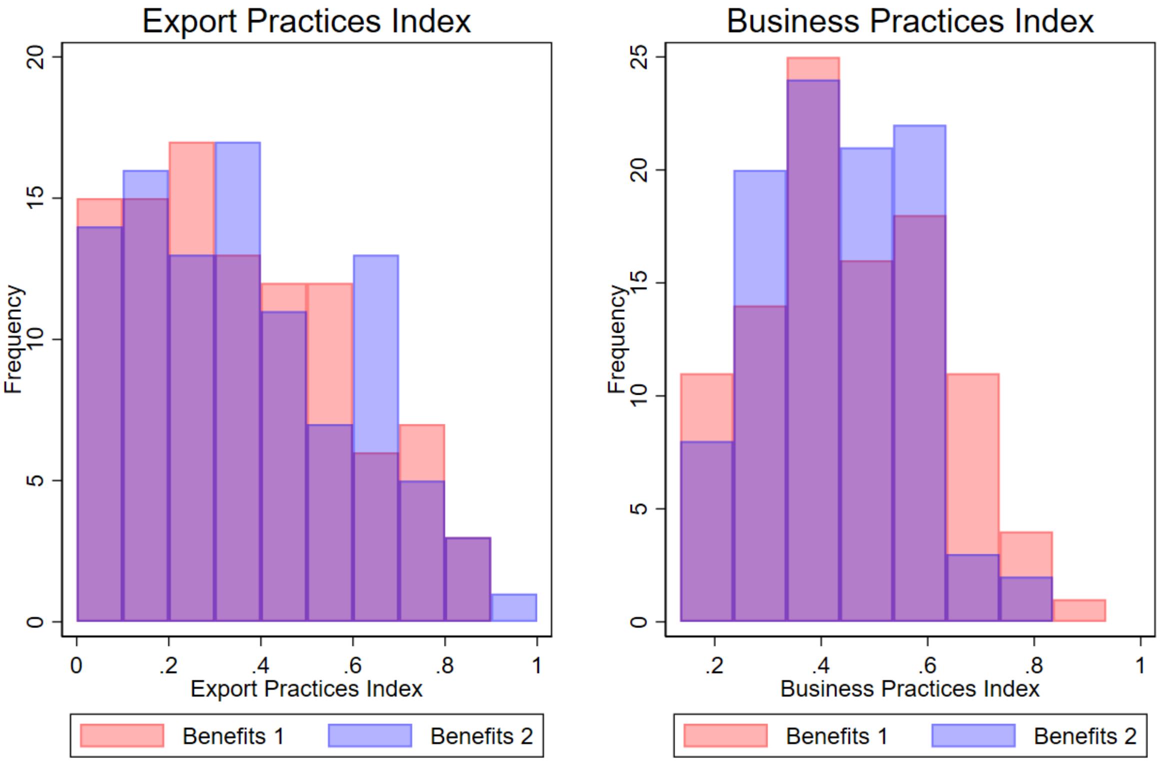


Figure 2: Distribution of Baseline Firm Outcomes

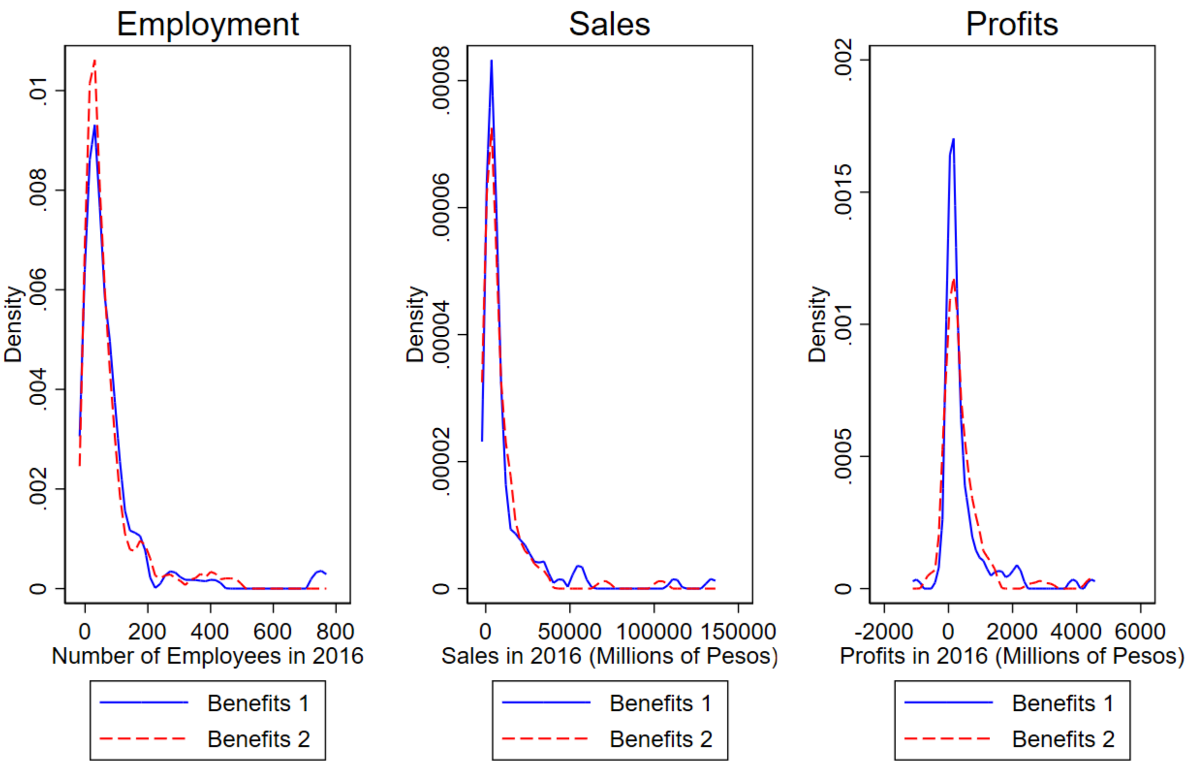
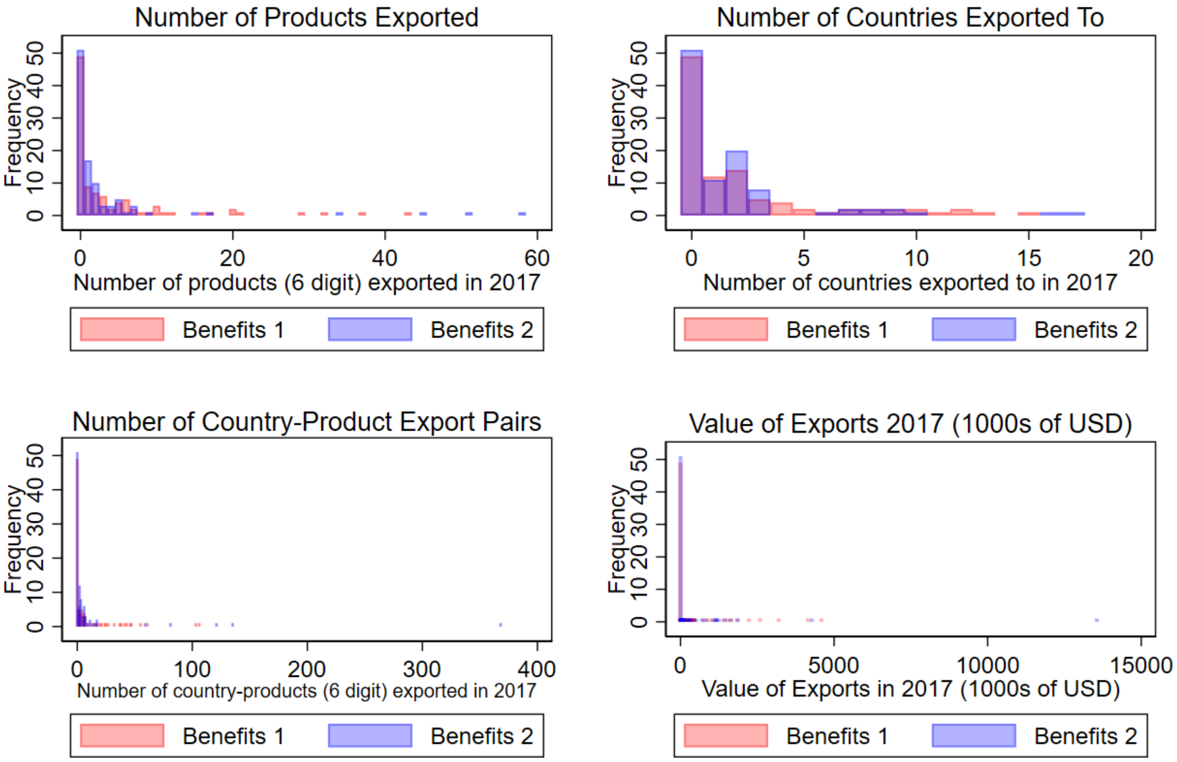


Figure 3: Distribution of Export Outcomes



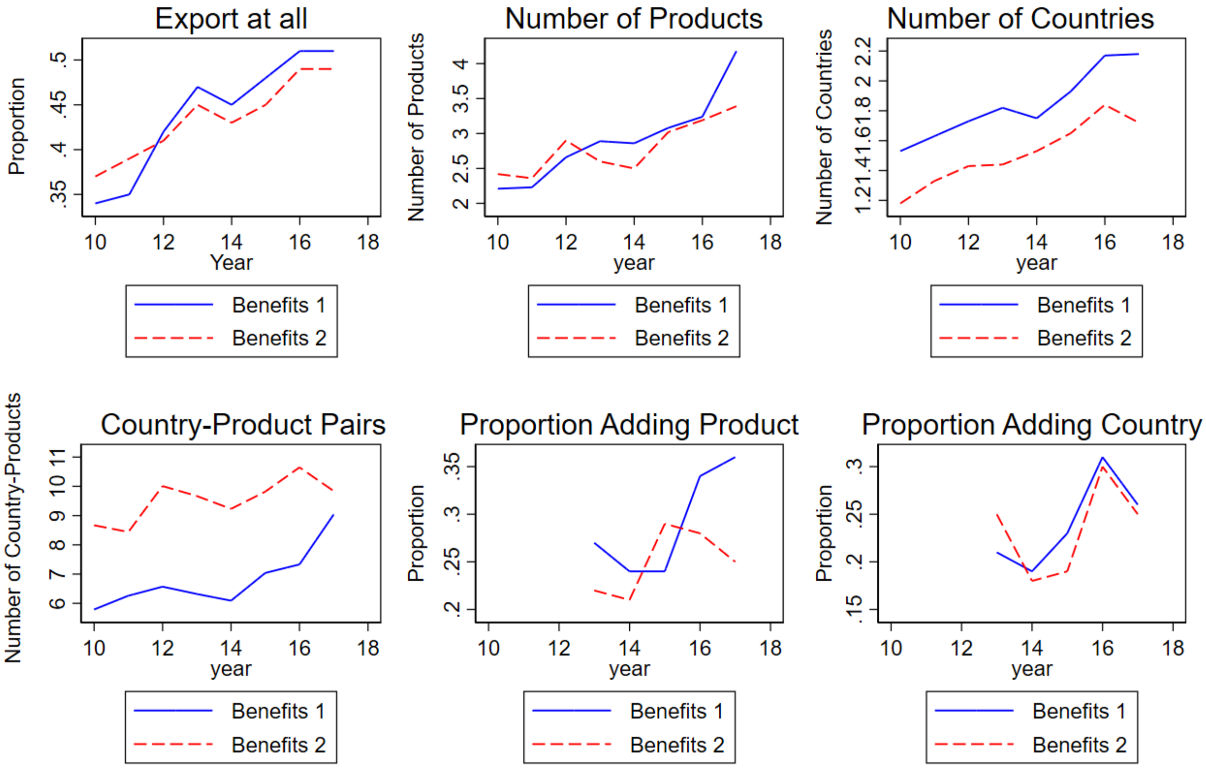
Number of Products Exported

Number of Countries Exported To

Number of Country-Product Export Pairs

Value of Exports 2017 (1000s of USD)

Figure 4: Export Means Over Time



Data collection and processing

- What are the key data sources? What data collection procedures and instruments will be used?
- What is the rule for terminating data collection (number of observations, available funds, available time, etc.)?
- How long will the data collection process take? If data will be collected at multiple points (longitudinal design), what is the proposed schedule (including enrollment, intervention delivery and outcome assessment)?
- What measures will you take to ensure data quality in terms of accuracy, consistency, bias, and completeness (e.g. double entry if manual entry of handwritten forms, audits of survey work, etc.)?

The key data sources are the following:

Export Administrative Data: using each firm’s national tax identification number (NIT), we can match them to official administrative data on export transactions. This database captures the product, country of destination, and value of export transaction, and is available at the monthly frequency. We aggregate this data up to the annual level. We have access to this data at the firm level, beginning in 2010. We will use 5 years of pre-intervention data, and then data for 2018, and 2019 for our initial analysis. Since this data is administrative data, firms can be tracked indefinitely, and we plan on subsequently tracking outcomes through at least 2020.

PILA database: the PILA is the platform through which all Colombian firms pay social security taxes for their employees. It therefore allows measurement of formal employment, at a monthly frequency. We will match firms to this data based on the NIT, and use up to 3 years of pre-intervention data, along with data for 2018 and 2019 for our initial analysis. As with the export administrative data, this data can be subsequently used to track firms over longer periods.

Firm Survey of Export and Business Practices: this will be taken at the end of the intervention, as an in-person firm survey of export practices. We expect this to be undertaken by Innovations for Poverty Action, Colombia. Interviewers will be blinded to treatment status to reduce bias in the collection of business practices. Standard data quality checks will be undertaken, including random callbacks, computerized error checks, and checks for completeness. Expected date for collection June-August 2019.

Annual Manufacturing Survey (EAM): this is an annual manufacturing survey collected by the Colombian Statistics agency (DANE). We will work with them to add the firms from our study that are not randomly selected as part of their sampling to be a booster sample. Data are collected once per year per firm, and have the advantage of having high response rates. These data are well-regarded for quality standards, taking particular care to measure key variables needed for productivity measurement such as input and output prices. However, as noted above, access to this data comes with a time lag, and must be done in a datalab, with only summary output released.

Variations from the intended sample size

- Do you anticipate any challenges in collecting data (attrition, non-compliance with the assigned treatment, etc.) and what measures do you plan to take to prevent them?

We will work closely with the PTP to monitor compliance with treatment assignment. We are confident that no firms allocated to the control group will end up receiving the intervention. The main concern is then whether there is sufficient take-up of the program, so that we have sufficient statistical power. As noted above, our power calculations are based on an anticipated treatment effect from the literature that comes from training program situations that have 65% take-up. We expect to have 80% take-up of the intervention amongst the treatment group, since firms have self-selected on interest in the program (unlike some of the interventions which do not screen on interest), and because consultants will be coming to visit the firms on-site, rather than requiring firms to travel (as is needed for training programs). Firms will receive additional messages encouraging participation from PTP and the World Bank.

Attrition is a key challenge in measuring firm outcomes. Our main solution to this challenge is to rely heavily on administrative data. Our primary outcome measures all come from export data that are available for all firms with no attrition. Likewise, the formal employment from the PILA will not have this attrition either. The EAM survey has high response rates due to the official statistical function of the survey, and we expect at most 10% of firms not to comply. The main attrition concern will then be in collecting business practice data from the firms through the survey. We will aim to minimize attrition through multiple callbacks, the use of prizes and incentives for firm response, and through minimizing the length of the questionnaire. We will test whether the attrition rate varies significantly with treatment status, and whether or not the sample responding to the survey remains balanced on baseline observables. If attrition varies significantly with treatment status, or different types of firms respond in treatment versus control, then we will use both Lee (2009) bounds and inverse-probability weighting to examine robustness of our impacts on business practices to this attrition.

Pilot data

- Summarize any pilot data used in preparation for this submission. These can be included to establish reality checks, effect size estimations, feasibility, or proof of principle.

Pre-intervention data are summarized in Tables 1 and 2, and Figures 1-4, and discussed in the context of out power calculations.

3. Empirical Analysis

Please use this section to present your strategy for statistical analysis. In the appendices section of this submission, please also include any computer programs, configuration files, or scripts which will be used to run the experiment and to analyze the data.

Statistical methods

- What statistical methods will be used to analyze the data and what are their underlying assumptions?
- How will the study deal with missing values?
- How do you define and handle outliers?

Frequentist Approach

Our frequentist approach will use an Ancova specification of linear regression to estimate the intention-to-treat effect. Our estimating equation for the ITT impact on outcome *y* of firm *i* being assigned to treatment (benefits 2) versus being assigned to control (benefits 1) takes the form:

$$y_{i,t} = \alpha + \beta Treat_i + \sum_{s=1}^{5} \gamma_s\, y_{i,t-s} + \sum_{s=1}^{5} \theta_s 1\left(missingy_{i,t-s}\right) + \sum_{j=1}^{54} \delta_j\, 1\left(i\epsilon strata_j\right) + \varepsilon_{i,t} \qquad (2)$$

Where $y_{i,t-s}$ is the *s*th pre-intervention lag of the outcome of interest, set to zero if this lag is missing, and $1\left(missingy_{i,t-s}\right)$ is a dummy for a missing value of the *s*th pre-intervention lag; δ_j are randomization strata fixed effects (following Bruhn and McKenzie, 2009); and β is the intent-to-treat effect. Robust (Eicker-White) standard errors will be used. While our main focus will be on estimating the ITT, we can also instrument receipt of the Benefits 2 (intervention) with random assignment to this group, to get the treatment effect on the treated (TOT). This requires the additional assumption that simply being selected for treatment has no effect on outcomes if firms do not participate in the intervention, which seems plausible ex ante. A linear model will be used for all outcomes, including a linear probability model for binary outcomes.

The key assumption underlying equation (2) is the stable unit treatment value assumption (SUTVA). This will be violated if treated firms compete directly for export sales with control firms, so that additional export success for the treated firms may come from competing away business from the controls.[9] The sectoral heterogeneity of firms in our sample helps in this regard. Using the 6-digit product code, 62% of firms do not have a single other firm in the study exporting the same product as them, 70% do not have any other firm exporting the same country-product combination, and 94% have 3 or fewer firms exporting the same country-product combination. Moreover, while some of the more common 6-digit product codes are very specific (e.g. uchuva (cape gooseberry), granadilla (yellow passionfruit), gulupa (purple passionfruit), pitaya (dragonfruit), and tomate de arbol (tamarillo), all types of tropical fruit)[10], others of the more common product categories have more within-category heterogeneity (e.g. cotton t-shirts, long and short trousers for women and children, shirts and blouses of artificial or synthetic fiber, miscellaneous plastic products, miscellaneous steel products). Our working assumption is therefore that any export growth from the treated firms is unlikely to be primarily business stealing from control firms. However, we will test this assumption by adding two additional variables to equation (2): the number of other firms in the study that were exporting the same product at baseline, and the number of these assigned to treatment. We can then test whether firm export outcomes vary with the number of other firms in the same product category assigned to treatment.

We have discussed our approach to missing data when defining our key outcomes and databases above – namely that the administrative database on export outcomes will not have missing data (a firm that does not appear will be assumed not to be directly exporting), and that we will test for whether attrition in collection of the business practice data are correlated with treatment status or causes imbalance in baseline observables, and use bounding and re-weighting approaches if this is the case.

Our approach to outliers is also contained in our definitions of the primary outcomes: namely we define there when we will winsorize, and which variables we will transform using the inverse hyperbolic sine transformation. As an additional check of robustness to outliers, we will re-estimate our impacts on primary outcomes by omitting the randomization strata with the 19 firms who were export outliers according to their baseline data.

Bayesian Approach

Our Bayesian approach will comprise both a Bayesian estimation of the same linear regressions proposed for the frequentist analysis on means and quantiles of outcome distributions, and in addition an estimation

[9] Another potential concern would be positive spillovers, if control firms learn business practices from treated firms. The fact that these firms are geographically and sectorally diverse, and are a small fraction of all firms in Colombia makes this unlikely, but we will ask in our follow-up business practices survey whether firms have learnt practices from other firms, and if so, the names of the firms they have learned from.

[10] Only 11 firms out of the 200 are in the fruit sector, but since fruits are exported to many countries, they are among the most common product codes in the export database.

of treatment effects via Bayesian imputation of the unobserved potential outcomes as in Imbens and Rubin (2015). These two approaches are outlined below. Because Bayesian analysis is less common in applied economics and best practices for inference in these settings are less well-established, this part of the pre-registration contains more allowances for unforeseen issues. In several cases we will fit 2-3 models and perform model selection to find the best, and we also describe what we will do in the event of unforeseen circumstances preventing us from following our plan. Part of the goal of this project is to generate greater understanding of what works in applied Bayesian econometrics, and provide new lessons for empirical economists about the relative strengths and weaknesses of Bayesian methods.

(1) Bayesian Linear Regression on Means and Quantiles

The Bayesian linear regression analysis will use the same conditional moment specification as equation (2) for both means and medians, but will perform inference by combining a likelihood with a prior via Bayes' Rule. In practice, the resulting posterior can be challenging or impossible to compute analytically and must be approximated using Markov Chain Monte Carlo Methods. We will use Hamiltonian Monte Carlo algorithm in which tuning parameters are automatically disciplined using the No-U-Turn-Sampler, implemented via the software package Stan. For a further discussion of this implementation see Meager (2016).

With a sample size of 200, the likelihood is typically parametrically specified due to concerns that a nonparametric method will overfit the data. Based on an assessment of the existing administrative data, we can pre-specify our chosen likelihoods for each variable. The 8 variables are:

1. *Extensive margin: Export at all in the past year:*
2. *Number of Distinct Products Exported in the past year:*
3. *Number of Different Countries Exported to in the past year.*
4. *Number of Distinct Product-Country Combinations Exported in the past year:*
5. *Export innovation (new product-country combination):*
6. *Total Export Value in the past year*.
7. *Export Labor Productivity:*
8. *A standardized export outcomes index*

They fall into several likelihood categories:

1. Binary Variables (1, 5): Bernoulli trial with a Gaussian link function and logit link function; aka a Bayesian logit and probit. Headline result will be chosen by Leave-One-Out-Cross-Validation.
2. Count Variables (2,3,4): Negative Binomial model (which nests Poisson regression as a special case).
3. Weakly positive variables with large numbers of zeroes (6,7): A mixture distribution of a spike and either a lognormal tail or a pareto tail. The sorting into the mixture components will be estimated as logit model.
4. Continuous variables (8): Gaussian model.

Notice that in order for the analysis on the binary outcomes to be more comparable across analysis types, we will also run the frequentist Logit/Probit and report the results of those for the purposes of comparison. In addition, while for weakly positive variables with many zeroes this spike and slab model is our preferred specification (in line with Meager (2017)), to better allow comparison with the frequentist model we will also transform to inverse hyperbolic sine and conduct a Bayesian analysis on the transformed data using a Gaussian likelihood.

In each case we will also apply the conditional moment regression specification on the other moments of the distributions. For example, lognormals have means and variances, and allowing treatment and covariates to affect the variance of the outcome is both informative in itself and enables us to capture heteroscedasticity. All models will be assessed for fit, and for their MCMC performance. In the unlikely case of intractably poor performance due to either nonconvergence or divergent transitions, alternative models or additional constraints on the existing models will be sought at that time.

However, given the large number of strata dummies, we can already anticipate that another adjustment to the specification is likely to be needed for Bayesian regressions. The current regression analysis proposes to fit a model with at least 66 parameters to 200 data points, which raises the possibility of overfitting and poor inferential properties in both a Frequentist and Bayesian setting. This issue is a higher priority to address in the Bayesian setting because all posterior inference is inherently joint. Hence, even in linear models the uncertainty tends to propagate unless the relevant covariates are truly orthogonal to one another in the finite sample. One typical solution in Bayesian inference is to reduce the number of effective parameters by placing a random effects structure on some of the parameters that are less central to the analysis. Such structure constrains the fit of the model and encourages those parameters to be zero unless the evidence for non-zero effects is strong (ie unless the nonzero coefficient fits the data much better). Applying a Gaussian structure to the distribution of these parameters regularizes the estimates exactly as does a Ridge regression from the Machine Learning literature.

Because it is unclear ex-ante how serious this potential problem will be, we will provide several analyses, including a random effects structure on all coefficients except the intercept and treatment effect, a tailored random effects structure based on inference conducted on the existing administrative data, and a completely unmodified structure with no reduction in effective parameters (this last option is likely to perform poorly, but useful to compute as a comparison). Since part of the goal of this exercise is to also generate practical guidance for other impact evaluations, this comparison will be useful in illustrating sensitivity to different decision choices.

The tailored random effects structure will form the headline result. We propose a model as follows. We will not constrain the role of the 3 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. But their interactions, which define the exact strata assignment and thus the dummies for strata, will be subject to a random effects structure centred at zero.

Bayesian quantile treatment effect analysis is a relatively new aspect of Bayesian inference, and our approach here is likely to require more adjustment than the above. Hence, we outline our general plan only. First, we will pursue the currently recommended implementation of Bayesian quantile estimation using the

Asymmetric Laplace Distribution (Yu and Moyeed, 2001). This can be implemented in Stan but we note that due to the lack of differentiability of certain parts of the likelihood, there is a risk that performance may be poor for these models. We will assess MCMC performance and convergence as recommended in the Stan Manual (2018) and seek alternative computational methods or alternative models and implementations if necessary. We will attempt to perform joint inference on all the quantiles given that we know their underlying relationship and constraints on these objects; this may require methodological innovation. We will be using diffuse, weakly informative priors on the quantile treatment effects centred around the prior on the mean effects. Note that here, as in the case of the frequentist regressions on quantiles, there is a greater risk of overfitting due to estimation of a greater number of parameters than in the mean models. We will attempt to prevent overfitting by conditioning only on the 5 dummies that define the main strata, and will not run the model with all 54 strata dummies.

(2) Bayesian Imputation of Potential Outcomes

Bayesian imputation of potential outcomes provides a complementary analysis to the above that permits a more generative modelling approach, which is both potentially more informative and as a result are often more computationally stable. This approach is Bayesian model-based imputation with covariates and is explained fully in section 8.7 in Chapter 8 of Imbens and Rubin (2015), but summarized here.

First we will specify the joint distribution of the potential outcomes $Y_i(0)$, $Y_i(1)$, for i = 1…N, conditional on covariates X including treatment assignment status (ITT) and conditional on unknown parameters θ. We envisage that this f (Y(0), Y(1)|X, θ) will typically be either bivariate Gaussian or Log-Gaussian, or Binomials with probabilities governed by bivariate Gaussian link functions (a multivariate probit). The parameters that govern this joint distribution will be those involved in the specification of the conditional mean, which will be specified such that covariates may enter, and the variance-covariance matrix. Note that these conditional mean specifications can allow the previous value of the outcome variable to enter as a covariate. Note further that we will allow the dependence of the potential outcomes on the covariates to differ across the two types of outcome.

Note that we will follow Imbens and Rubin (2015) in assuming that ($Y_i(0)$, $Y_i(1)$, X_i) are conditionally exchangeable given the parameters and hence can be treated as i.i.d. due to De Finetti’s theorem. Because we can factor the trivariate distribution and impose a prior on the generating parameters for X_i which also factors out from the rest of the prior, it is valid to consider only the model f (Y(0), Y(1)|X, θ).

Using this functional form assumption, we will be able to derive the conditional likelihood distribution of the unobserved (missing) potential outcomes. Then we will place priors on these unknown parameters and combine them with the joint likelihood from step 1. The choice of priors is discussed in the section below. However we do note that Imbens and Rubin (2015) suggests that in randomized experiments the choice of priors may have little influence on the results. This is something that we will investigate in our setting.

We will then derive the full joint posterior distribution of all the unknown parameters using MCMC methods. Finally we will compute the posterior distribution of the estimands of interest. This will include the average treatment effect as well as the full distribution of treatment effects ($Y_i(1)$ - $Y_i(0)$), which can be represented by the quantiles of that distribution.

An issue raised in Imbens and Rubin (2015) in section 8.6 is the fact that by definition the data cannot provide information about the correlation between the Y(0)s and Y(1)s, because we never observe both of them for any unit. However, this correlation (denoted ρ) does influence the uncertainty one should have about the distribution of treatment effects. While Imbens and Rubin note that the influence is often small in large samples, they do suggest that in the absence of true prior information it might be desirable to conduct conservative inference, either by choosing the ρ that maximizes posterior uncertainty or by assuming independence. While we will elicit true prior information about the value of ρ and use that information, we will also conduct inference assuming perfect independence and perfect correlations in both directions.

Ex-post we will also assess the extent to which overfitting is likely to have occurred in both frequentist and Bayesian statistical analyses, by calculating the leverage of the analyses (see Young 2016 and by leave-one-out cross validation performance.

Both Bayesian approaches use priors to regularize estimates and reduce the influence of outliers, rather than trimming or winsorizing the dataset. In the event that the Bayesian results differ substantially from the frequentist results it may be ex-post of interest to re-run the Bayesian analysis on the winsorized data to isolate the source of the difference. Finally we note also that the estimation of quantile treatment effects is inherently less sensitive to extreme observations than mean effects.

Prior Elicitation and Specification

The Bayesian methods above both require the specification of priors on the unknown parameters. In many cases, Bayesian analyses can proceed using priors from the previous literature or, in the absence of this, simply using weakly informative priors (Gelman et al, 2008).

However, we will be able to elicit prior information from stakeholders in our study including the firms themselves. Note that it is neither feasible nor necessary to elicit prior information on every single unknown parameter from these sources for our study, because there is baseline / administrative data that can inform the priors on most of the parameters, with the obvious exception of the treatment effect. Due to time constraints and taking into consideration the parameters over which experts are likely to have well-defined prior beliefs, we will be eliciting information on only a subset of unknown parameters.

We plan to elicit priors from 3 separate groups of knowledgeable individuals: firms themselves, policymakers and academics. We aim to elicit beliefs from at least 5 persons in each category. This will be done using the “balls and bins” approach to eliciting subjective expectations, discussed in Delavande et al. (2011). Notice that since our main specification uses ITT as the treatment variable, we need prior information on the ITT.

The parameters on which we plan to elicit prior information are:

- The β parameter for each of the 8 key outcomes which captures the average treatment effect in the regressions.
- The correlation coefficient between the potential outcomes under treatment and no treatment for the Imbens and Rubin (2015) method.

The literature-based priors will be determined based on the existing data of these firms before the intervention takes place, specifically from the relationships shown in table 3. Panel A of table 3 shows the correlation between the Export Practices Index and our various outcomes of interest. As noted in our power calculations, McKenzie and Woodruff (2014) find that standard business training interventions tend to increase business practices by 0.05 to 0.10. Since our intervention is more intensive than most in the literature, we use 0.10, and center our literature-informed priors around the assumed effects used in Table 4 for the power calculations. Uncertainty about how much the intervention will alter the EP index, as well as uncertainty in the extent to which we can extrapolate from the previous data, creates prior uncertainty here.

In addition to reporting the posterior inference for each of these four priors, we will combine these posteriors into a single inferential framework using the stacking procedure outlined in Yao et al (2018). We expect it will also be instructive to combine all the elicited prior beliefs into a single prior and perform a single inference.

Finally, in the interests of understanding Bayesian statistical practice, we will fit the models with the typical "default" priors that Bayesian analyses often use in the absence of elicited or reliable ex-ante information about the parameters. For example, Vivalt (2017) uses uninformative priors and Meager (2016) uses weakly informative priors typically centered at zero with dispersion 2-5 times greater than the observed dispersion in the outcome being studied. These choices are often made in order to reflect a wide range of beliefs on the unknowns, but a desire to regularize parameter estimates towards zero to mirror common machine learning procedures. They also reduce the incidence of overfitting producing misleading inference far from zero. Hence, we will also compute the posterior given weakly informative priors on the treatment effects centered at zero with dispersion 5 times greater than the outcome's dispersion in the control group, or equivalently weak for the scale specified.

Statistical model

Provide the model in its *functional* form and submit math equations as editable text and not as images.

Multiple outcome and multiple hypothesis testing

- How will the study address false positives from multiple hypothesis testing?
    - If you plan to adjust your standard errors, what adjustment procedure will you use? (e.g., Family Wise Error Rate, False Discovery Rates, etc.)
    - If you plan to aggregate multiple variables into an index, which variables will you aggregate and how?

We will use two approaches to multiple hypothesis testing. The first is to define an index measure of our primary hypothesis outcomes (outcome 8 amongst our primary hypotheses). This aggregate index of standardized z-scores will serve as a summary measure of all export outcomes. Secondly, we have kept the list of pre-specified outcomes short within each hypothesis family, and then will calculate sharpened q-values that hold constant the false discovery rate following Anderson (2008).

Heterogeneous Effects

- Which groups do you anticipate will display heterogeneous effects? What leads you to anticipate these effects?

Given the heterogeneity in our sample, there are many dimensions along which we might think the treatment has heterogeneous effects (existing export behavior, firm size, existing state of export practices, sector, etc.). However, with our limited sample size, statistical power for examining treatment heterogeneity in multiple dimensions will be low. We therefore tie our hands by restricting ourselves to examining heterogeneity with regard to a single firm characteristic: *whether or not the firm had exported in the past three years*. Recall this is one of our randomization strata variables, and 58 percent of firms have exported in the past three years. This choice is motivated by the existing literature (both Kim et al. (2016) and Breinlich et al. (2017) find that existing exporters respond more to their information treatments); by the possibility that non-exporters face additional constraints that may prevent them converting better business practices into exporting; and by the policy interest in whether such a program works better at the extensive margin of encouraging firms to start exporting, or the intensive margin of helping existing exporters to export more. To estimate the heterogeneity of impact with this characteristic, we add an interaction between treatment assignment and having exported in the past three years to equation (2).[11]

A second approach we will use to examine heterogeneity in treatment impact is through estimation of quantile treatment effects for our continuous primary export outcomes. This will enable us to go beyond means and estimate impacts at different parts of the distribution. It is common for quantile treatment estimation not to converge when a large number of strata dummies are included. Therefore we propose to simply control for the stratification variables (exporting or not, medium firm, large firm, export outlier, and export practices index), rather than the full set of interactions that form the strata dummies. For the frequentist estimation, we will estimate the treatment effect at each 5th quantile from 5 to 95, and plot the estimated treatment effects and confidence intervals.

Our Bayesian approach to impact evaluation likewise will enable us to obtain quantiles of the posterior distribution, in addition to the posterior mean and standard deviation. In particular, the Imbens and Rubin method permits direct inference on the distribution of the individual treatment effects under the assumptions of the model. Even when Gaussian likelihoods are used in these models, uncertainty about the covariation can lead to non-Gaussian posteriors and thus will be able to capture nuanced differences in potential outcomes beyond simple location-scale shifts.

4. List of References

Ackerberg, Daniel, Kevin Caves and Garth Frazer (2015) “Identification Properties of Recent Production Function Estimators”, *Econometrica* 83(6): 2411-51.

[11] Note that the stratification dummies already capture the level effect of having exported in the past three years.

Anderson, Michael (2008), "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects", *Journal of the American Statistical Association*, 103(484), 1481-1495

Atkin, David, Amit Khandewal and Adam Osman "Exporting and Firm Performance: Evidence from a Randomized Experiment", *Quarterly Journal of Economics* 132(2): 551-615

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie and John Roberts (2013) "Does Management Matter? Evidence from India", *Quarterly Journal of Economics*, 128(1): 1-51

Bloom, Nicholas, Raffaella Sadun and John Van Reenen (2016) "Management as a Technology", Mimeo. Stanford University.

Bloom, Nicholas, Kalina Manova, John Van Reenen, Stephen Sun and Zhihong Yu (2017) "Managing Trade: Evidence from China and the U.S.", Mimeo. Stanford University.

Bloom, Nicholas and John Van Reenen (2007) ""Measuring and Explaining Management Practices Across Firms and Countries", *Quarterly Journal of Economics* 112(4): 1351-1408

Breinlich, Holger, David Donaldson, Patrick Nolen and Greg Wright (2017) "Information, Perceptions and Exporting – Evidence from a Randomized Controlled Trial", Mimeo. University of Nottingham

Bruhn, Miriam, Dean Karlan, and Antoinette Schoar (2018), "The Impact of Consulting Services on Small and Medium Enterprises: Evidence from a Randomized Trial in Mexico" *Journal of Political Economy*, 126(2), 635-687

Bruhn, Miriam and David McKenzie (2009) "In Pursuit of Balance: Randomization in Practice in Development Field Experiments", *American Economic Journal: Applied Economics*, 1(4): 200-32

Cadot, Olivier, Ana Fernandes, Julien Gourdon and Aaditya Mattoo (2015) "Are the benefits of export support durable? Evidence from Tunisia", *Journal of International Economics* 97(2): 310-24.

Delavande, Adeline, Xavier Gine and David McKenzie (2011) "Measuring subjective expectations in developing countries: A critical review and new evidence", *Journal of Development Economics* 94: 151-63.

Dellavigna, Stefano and Devin Pope (2017). "Predicting Experimental Results: Who Knows What?", *Journal of Political Economy*, forthcoming.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. "A weakly informative default prior distribution for logistic and other regression models." *The Annals of Applied Statistics* 2, no. 4 (2008): 1360-1383.

Girma, Sourafel, Yundan Gong, Holger Görg, and Zhihong Yu (2009). “Can Production Subsidies Explain China’s Export Performance? Evidence from Firm-level Data”; *Scandinavian Journal of Economics* 111(4): 863-891.

Groh, Matthew, Nandini Krishnan, David McKenzie, and Tara 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 (9): 1-23.

Gronau, Quentin, Alexander Ly, and Eric-Jan Wagenmakers (2018) “Informed Bayesian T-tests”, arXiv:1704.02479 [stat.ME]

Higuchi, Yuki, Vu Hoang Nam, and Tetsushi Sonobe (2015) “Sustained impacts of Kaizen training”, *Journal of Economic Behavior and Organization* 120: 189-206.

Hirschleifer, Sarojini, David McKenzie, Rita Almeida and Cristobal Ridao-Cano (2016). “The Impact of Vocational Training for the Unemployed: Experimental Evidence from Turkey”, *Economic Journal*, 126(597), 2115-2146.

Imbens, Guido and Donald Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An introduction*. Cambridge University Press, New York, NY.

Kim, Yu Ri, Daichi Shimamoto, Petr Matous and Yasuyuki Todo (2016) “Are seminars for export promotion effective? Evidence from a randomized trial”, Mimeo. University of Tokyo.

McKenzie, David (2011) “How can we learn whether firm policies are working in Africa? Challenges (and solutions?) for experiments and structural models”, *Journal of African Economics*, 20(4): 600-25.

McKenzie, David (2012) “Beyond Baseline and Follow-up: The Case for more T in Experiments”, *Journal of Development Economics*, 99(2): 210-21

McKenzie, David, and Christopher Woodruff (2014). “What Are We Learning from Business Training Evaluations around the Developing World?”, *World Bank Research Observer*, 29(1), 48-82

McKenzie, David and Christopher Woodruff (2017) “Business Practices in Small Firms in Developing Countries”, *Management Science*, 63(9): 2967-81

Meager, Rachael (2017). “Aggregating distributional treatment effects: A bayesian hierarchical analysis of the microcredit literature.” Working Paper.

Meager, Rachael (2018) "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments" Accepted at *the American Economic Journal: Applied Economics* in January 2018.

Vivalt, Eva (2017) “How much can impact evaluations inform policy decisions?”, Mimeo. Australian National University.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman (2018). "Using stacking to average Bayesian predictive distributions." *Bayesian Analysis* (2018).

Young, A. (2016). Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *London School of Economics, Working Paper,* Feb 2016

5. Appendices

If there is more than one appendix, they should be identified as A, B, etc. Formulae and equations in appendices should be given separate numbering: Eq. (A.1), Eq. (A.2), etc.; in a subsequent appendix, Eq. (B.1) and so on. Similarly for tables and figures: Table A.1; Fig. A.1, etc.

Appendix A: Definition of Export Practices Index

The Export Practices Index is the Proportion of the following 11 export practices that firms indicate they have in place at the time of application:

1. They have participated in trade fairs
2. They segment clients by international location
3. They travel to selected markets to understand consumers
4. They get distributor information through commercial missions
5. They get distributor information through trade fairs
6. They get distributor information through the export promotion agency
7. They get distributor information through advertising in destination countries
8. They participate in missions or public agency offerings to learn quality
9. They plan resources or training needed for export processes monthly
10. They plan production with time for external markets
11. They have quality certification

6. Administrative information *(required)*

Funding: Please list funding sources in this standard way to facilitate compliance to funder's requirements:

This work is currently supported by the World Bank under the CIIP Trust Fund.

(e.g. “This work was supported by the National Institutes of Health [grant numbers xxxx, yyyy]; the Bill & Melinda Gates Foundation, Seattle, WA [grant number zzzz]; and the United States Institutes of Peace [grant number aaaa]”).

If no funding has been provided for the research, please include the following sentence: “This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.”

Elsevier has established a number of agreements with funding bodies which allow authors to comply with their funder's open access policies. Some funding bodies will reimburse the author for the Open Access Publication Fee. Details of existing agreements are available here.

Institutional Review Board (ethics approval): If applicable, please include a statement confirming that all necessary ethics approvals are in place.

Declaration of interest: Please provide a statement on competing interests, even if you have no competing interests to declare.

Examples of potential conflicts of interest include employment, consultancies, stock ownership, honoraria, paid expert testimony, patent applications/registrations, and grants or other funding. If there are no conflicts of interest then please state ‘none’. You can learn more about our policies on conflicts of interests here.

Acknowledgments: Please list here individuals who provided help during the research, however are not considered authors (e.g., providing language help, writing assistance or proof reading the article, etc.).