Improving Management with Individual and Group-Based Consulting: Results from a Randomized Experiment in Colombia[*]

Leonardo Iacovone, *World Bank*

William Maloney, *World Bank*

David McKenzie, *World Bank*

Abstract

Differences in management quality are an important contributor to productivity differences across countries. A key question is how to best improve poor management in developing countries. We test two different approaches to improving management in Colombian auto parts firms. The first uses intensive and expensive one-on-one consulting, while the second draws on agricultural extension approaches to provide consulting to small groups of firms at approximately one-third the cost of the individual approach. Both approaches lead to improvements in management practices of a similar magnitude (8-10 percentage points). The group-based intervention leads to significant increases in firm sales, profits and labor productivity, while the impacts on firm performance are smaller in magnitude and less robust from the individual consulting. The results point to the potential of group-based approaches as a pathway to scaling up management improvements.

Keywords: Management, Employment, Scaling-Up Interventions, Colombia.

JEL codes: O14, O32, L2, M2

[*] The authors gratefully acknowledge the collaboration of Paula Toro Santana, the staff of CNP and DNP in Colombia, and project management and research assistance provided by Yesica Fernández, Cosma Gabaglio, Camilo Andrés Gutiérrez Silva, Pablo Villar, María Aránzazu Rodríguez Uribe and Innovations for Poverty Action Colombia. Funding is gratefully acknowledged from the DIME i2i Trust Fund, the Knowledge for Change Program (KCP), the World Bank and the IPA SME Initiative, as well as intervention funding from SENA. This study is registered in the AEA RCT registry AEARCTR-0000528. Since no identifying information was collected on human subjects, the study was exempted from the Innovations for Poverty Action IRB. Comments from the editor, five anonymous referees, Miriam Bruhn, Jasmin Chakeri, Patricio Dalton, Siddharth Sharma and participants in various seminars are greatly appreciated.

1. Introduction

There are large differences in the management practices used by firms within and across countries (Bloom and Van Reenen, 2007; McKenzie and Woodruff, 2017). These differences are strongly correlated with productivity, with Bloom et al. (2016) estimating that differences in management can account for 30 percent of cross-country productivity differences. An experiment with 17 textile firms in India provides a proof-of-concept that intensive individualized consulting can deliver lasting improvements in the practices of badly managed firms, resulting in productivity improvements of 17 percent (Bloom et al, 2013; Bloom et al, 2020). However, the intervention was implemented by an international consulting company under close supervision from researchers, and had a market value of $250,000 per treated firm.[1] This high cost is likely to be prohibitive for many small and medium enterprises (SMEs) to finance themselves, and for governments seeking to scale-up this to assist large numbers of firms.

This paper seeks to test two approaches that governments can use to scale-up management improvements. The first is to use a very similar intervention of intensive individualized consulting, but to use local teams of consultants to deliver the intervention at a lower cost of approximately $30,000 per firm. The second, more novel, intervention is a group-based approach that aims to deliver improvements at a reduced cost (around $10,000 per firm), and to leverage group-learning dynamics. We partner with the Colombian Government to conduct an experiment to measure the impact of these two competing interventions on SMEs in the Colombian auto parts manufacturing sector. Our sample of 159 firms with an average size of 58 employees, randomized into three groups of 53 firms, is an order of magnitude larger than that used in Bloom et al. (2013) and enables us to measure the impact of such a program when implemented at a multi-million-dollar scale by a government.

We show that the Colombian auto parts sector has similar levels of management practices to start with as the average Colombian manufacturing firm, which is low by global standards and similar to that in countries like India and Kenya with lower per-capita incomes. Both the individual and group-based interventions lead to improvements in management of similar magnitudes of 8 to 10 percentage points (relative to a control mean of 56 percent of structured managerial practices being

[1] The authors report that they paid an academically discounted rate of $75,000 per firm, with the consulting firm estimating a market price of up to $250,000 for those services.

implemented). This improvement is broad-based, with improvements in just over half of a detailed set of 141 practices measured. We then link firms to administrative data on employment and to a government annual panel survey to track firms for 3 to 4 years post-implementation. We find that the group-based interventions has grown the treated firms, with a statistically significant 6 to 15 worker increase in employment relative to the control group; a 28 to 33 percent growth in sales and production, a 5 to 26 percent increase in profits, and a 43 percent increase in value-added. Labor productivity increases by 11 to 14 percentage points, although this impact is not statistically significant. The impacts of the individual treatment on firm performance are smaller in magnitude and more sensitive to functional form and sample composition. Employment increases 2 to 7 workers, and sales a statistically insignificant 5 to 13 percent. The group-based intervention clearly dominates the individual intervention on a cost-benefit basis.

This work contributes to at least three literatures. The first is a general literature on improving business practices and management in firms. Most of this literature has focused on short training courses and microenterprises (see McKenzie 2020 for a recent review). However, several studies show the potential of more intensive individualized consulting to improve management in small and medium enterprises. In addition to Bloom et al. (2013)'s work in India, this includes Bruhn et al. (2018) with firms averaging 14 workers in Mexico, and Higuchi et al. (2017) with firms averaging 20 workers in Vietnam.[2]

Second, the paper contributes to an emerging literature on interfirm interactions and social learning that has highlighted the ability of firms to improve their business practices when formed into groups or paired with other firms that can serve as role models (e.g. Cai and Szeidl, 2018, Chatterji et al. 2018, Dalton et al. 2018, Lafortune et al. 2018). Whereas standard consulting transfers general knowledge, working in groups can bring additional gains: improved information due to negotiating common context-specific problems (Ray 2006, Brooks et al. 2018); or validation that suggested practices are, in fact, useful in the local context, and learning as a group about how best they can be modified. Working in groups or networks may also facilitate better matching of suppliers (Cai and Szeidl, 2018). A key distinguishing feature of our set-up is that we are working with SMEs that are more complex organizations than the microenterprises or small farms that have

[2] A related quasi-experiment provided evidence of the long-term impact of participating to the Productivity Program which allowed Italian firms to participate to study trips in US plants followed by consulting sessions of US experts at Italian firms (Giorcelli, 2019).

been the focus of the development literature. Our group intervention therefore not only has firm owners talking with other firm owners, but specialized staff from each firm sharing experiences and learning from their counterparts in other firms. The result is a much more intensive interaction at multiple points in the organization. It relates to the insight of Chandler (1977), that sharing of new technological and managerial lessons among a professionalized managerial class was a critical component of the emergence of sophisticated organization structures in large firms in U.S. development.

Our interviews confirm the finding of Bloom et al (2013) that many practices are not adopted because lack of knowledge about the practice or firm owners not thinking they were worth implementing, suggesting that collective learning about them may have important additional effects. In fact, we find evidence that the adoption of new practices arising through working in groups is not due to receiving information from the best in the group serving as a kind of local consultant. Rather, improvement is correlated with the overall learning of the group which suggests some coordinated experimentation and learning. The group intervention, then, is potentially more than a way of saving money, it may foster additional learning dynamics that lead to more effective adoption of practices.

Finally, this paper contributes to a broader literature on how to scale-up policies from promising researcher pilot studies (e.g. Banerjee et al, 2017, Bold et al, 2018). Government implemented programs tend to have smaller effect sizes than those implemented by researchers or NGOs (Vivalt, 2019), which, coupled with the high cost of Bloom et al. (2013)'s individual consulting intervention, raised questions as to their proof of existence that bad management can be improved could then provide a model that could actually be used by governments. Our results show the promise of group-based consulting as a pathway to greater scale by lowering the cost of delivery and delivering improvements in management and growth in scale for participating firms.

2. Context and Sample

2.1 Choosing the Industry and Sample

Labor productivity in Colombia is low, with it taking around four Colombian workers to produce what one worker does in the United States (Londoño, 2017). As a result, improving productivity is a priority for government policy. The Government of Colombia was interested in testing whether the productivity improvements from better management demonstrated in India by Bloom et al.

(2013) could be achieved at a larger scale in Colombia, as well as in generating more employment in these larger firms. In order to test different approaches, they wanted to choose a sector that was thought to have sufficient numbers of firms, to have production in a number of locations throughout the country, was thought to have some potential for growth, and was thought to be similar enough to other industrial sectors in the country that the results from this pilot could be applicable to other industries. These criterion led to the selection of the auto-parts sector. This sector consists largely of second-tier suppliers to large car manufacturers, producing parts like fenders, tires, suspension parts, plastic parts, paints, etc. that are sold to the assemblers that supply directly national and international car manufacturers as well as to retailers of spare parts. Appendix 1 provides some examples of the products. The auto parts sector in Colombia employs approximately 25,000 workers, and sells both to car and bus manufacturers within Colombia, as well as exporting approximately $US500 million each year, with Ecuador, Venezuela and Brazil the main export markets (Proexport Colombia, 2012).

Public announcement of the *Programa de Extensionismo Tecnológico* (Technological extension program) was made in April 2012 (Appendix 2 contains the full timeline), and firms were also informed of the program through the car manufacturers such as Sofasa (which assembles Renault cars in Colombia), General Motors, and Busscar (which manufacturers buses). To be eligible firms had to be legally registered, in business for at least two years, be a first or second-tier supplier to the automobile industry, and be located in one of four regions where the program focused: the departments of Antioquia, Cundinamarca, Valle del Cauca, and the Eje Cafetero (Coffee Region which comprises the departments of Caldas, Quindío, Risaralda and Tolima). The firms were told the program would offer assistance in improving production practices in order to improve profitability, productivity and competitiveness, and that the program would not require any payment by the firms, but that they would need to commit time and effort of their workforce to supply information required and to implement suggestions made.

Public provision of the program to firms was justified both with reference to the overall policy objective of improving productivity, as well as due to the presence of several market failures that prevent firms from improving management on their own. A first issue is that of information: many badly managed firms do not know they are badly managed, with data from the World Management Survey showing that Colombian managers perceive their firms to be much better managed than

the reality.[3] Secondly, even if firms know they need to improve, they may be unable to identify which providers can offer good services, may lack the financial resources to pay for consulting, and a lack of insurance may prevent them from investing in an activity with uncertain payoffs.

218 firms applied for the program. 180 of these were accepted in the preliminary step, with the remainder rejected for being too small, or for only being distributors rather than manufacturers of parts. 11 firms then dropped out, so 169 firms formed the group to take part in the first, diagnostic, phase of the project. Following the diagnostic, we dropped firms with fewer than 10 workers, to leave a sample of 159 firms for the experiment. Appendix 3 uses data from the annual Colombian manufacturing survey (EAM) to calculate the proportion of firms in the industry that participated in the program, and to compare the characteristics of the non-participants to our experimental sample. Our firms cover a wide range of product categories, including metal, glass, plastic, and rubber parts, and so are only 3.3 percent of all Colombian firms in these broad categories. If we consider the subset of firms in ISIC code 2930 (manufacture of parts, pieces, and accessories for automobiles), then our experimental sample contains 20 percent of all Colombian firms in this sector, and 34% of those with 50 to 250 workers. Average size and productivity for participants are similar to those of non-participants, but the distributions are different, with participants being drawn more from the middle of the size distribution, and very small and very large firms non-participating.

2.2 Random assignment and firm characteristics

Firms were randomly assigned to three groups of 53 firms each. Since the number of firms in each group would be small, we aimed to improve balance on observables by forming matched triplets of firms, choosing this grouping in a way to minimize the Mahalanobis distance between firms in a triplet in terms of their geographic location, size, labor productivity, and management practices.[4] This took place in November 2013, after the diagnostic phase (described below). Then within each triplet, firms were randomly allocated to a control group and two treatment groups: an individual-consulting treatment group and a group-consulting treatment group.

[3] Colombian firms had an average WMS score of 2.50 in 2014 (described below), but an average perceived score of 3.76. In contrast, U.S. firms had an average WMS score of 3.32, and perceived score of 3.57.

[4] Location consisted of Cundinamarca and Valle regional dummies; firm size consisted of dummies for small (10 to 50 workers) and medium size (51 to 310 workers), as well as for the number of employees; management practices consisted of indices for practices in human resources, production, logistics, marketing and finance; as well as for seven individual management practices identified as priority areas in many diagnostic plans.

Table 1 provides some summary characteristics of the firms, along with their means by treatment group status. The mean (median) firm has been in business for 24 (23.5) years, with only 20 percent having been in business for fewer than 10 years. A key feature of the data is that firms are heterogeneous in terms of size and product produced. Firms had a mean of 59 and median of 40 employees at the time of application, with 59 percent of the firms classified as small (10-50 workers), and the remainder as medium (51 or more workers), with the maximum being 310, and the 10-90 range being from 13 to 119 workers. Mean sales were approximately USD2.3 million in 2013, with a 10^{th} percentile of USD240,000 and 90^{th} percentile of USD5.3 million showing the large variation in firm size. These are almost all single plant firms, with the main subsectors being metal products (60%) and plastic products (18%). The sample also includes firms making rubber products (5%), chemical products such as injection molding (4%), electronic components (4%), as well as firms working with leather, wood, and glass. 94 percent are tier 2 firms in the value chain, with 6 percent tier 1.[5] Tier 2 firms do not supply parts directly to automobile manufacturers, but instead supply products to a wide range of clients including Tier 1 suppliers. The median firm reports having 50 regular clients in Colombia at baseline, and only 14 percent have fewer than 10 regular clients. Forty-five percent of firms had exported in at least one month of 2013. Half the firms are located in the Cundinamarca region, which includes Bogota, with the region of Valle del Cauca, which includes Cali, the next biggest.

Management practices were measured in terms of 141 individual practices based on an assessment developed by the Colombian National Productivity Center, and classified into five areas: financial practices (made up of 29 individual practices), human resource practices (20), logistics practices (31), marketing practices (22), and production practices (39). Each practice was scored on a five-point scale, where 1 indicates that the practice is not used, and 5 that it is implemented and under control. Scores were then aggregated and calculated as a percentage of the maximum possible score for that index. Appendix 4 provides more details of the specific measures. At baseline average scores for these practices range from 43 (human resources) to 51 (financing practices), relative to a potential maximum score of 100, indicating that firms have significant room to

[5] Tier 1 means that the firm directly supplies the original equipment manufacturer (e.g. Ford, Suzuki, etc.), while tier 2 means the firm supplies a tier 1 supplier without supplying the vehicle manufacturer directly.

improve on these practices. We refer to this as the Anexo K (Annex K) management practices measure, with this terminology referring to the form used to collect this data.

Table 1 shows that while the random assignment was able to achieve balance on most baseline variables, there are a couple of imbalances when test equality of means a variable at a time. These reflect the difficulty of balancing many variables in a relatively small sample of heterogeneous firms. For example, the control group is more likely to be in metal products than either treatment group and starts with lower labor productivity. However, our overall omnibus tests of joint orthogonality cannot reject that these variables do not jointly predict treatment status. In our analysis we use firm fixed effects or controls for the baseline value of interest to make the firms more comparable and reduce the effect of this heterogeneity.

2.3 External validity and comparison to Bloom Van Reenen Management Practices

In 2013, prior to the interventions, we commissioned the LSE survey team responsible for the Bloom and Van Reenen (2007) World Management Surveys (WMS) to apply their methodology to a random sample of 180 firms representative of the Colombian manufacturing sector, as well as to a sub-sample of 72 companies from our sample with 40 or more employees.[6] Appendix 7 summarizes this survey process, and provides three key results. First, the mean and distribution of WMS management practices scores for our auto parts firms is similar to that of the overall manufacturing sector in Colombia (2.38 versus 2.54). Second, Colombia's average management practices score shows firms are, on average, poorly managed in global terms, but similar to many other developing countries. The average score is just below that of firms in India and just above that in Kenya in the WMS. The auto parts sector in Colombia is thus a fairly typical sector for both the country, and for developing countries as a whole, in terms of management practices.

A final use of this baseline WMS data is to compare the Anexo K management measure, our main measure of management used in this paper, to the WMS. Appendix 5 shows that the two are significantly correlated in the cross-section at baseline, with a correlation of 0.26 between the two overall indices. The Anexo K has a stronger correlation (0.44) with the monitoring subcomponent of the WMS, reflecting a particular emphasis on measurement and monitoring than on other management practices.

[6] This size restriction was made since the WMS was designed for firms with 50 or more employees.

2.4 Macroeconomic context

The Colombian auto parts sector had sales grow at an annual average of 5.4 percent per year over the 2002 to 2012 period leading up to our experiment (Reina et al, 2014).[7] At the start of our study, imports averaged 68 percent of total sales in the sector, and were the main source of competition for most firms in our study. However, the country was hit by a combination of external and internal shocks starting in late 2013, which resulted in a large depreciation of the peso, from an average of 1930 COP to the USD in 2013 to approximately 3000 COP to the USD in each of 2015, 2016, and 2017. Domestic new vehicle sales fell from 326,000 units in 2014 to 238,000 units in 2017, a 27% drop (BBVA Research). Export sales of auto parts fell 51 percent in dollar terms over the 2013-2016 period, driven by weak economies in the main export destinations of Venezuela, Ecuador and Brazil. The aggregate context is thus one of weakening overall demand for the sector, but where the weakened currency increased competitiveness against imports. Real sales of domestic production were then fairly flat over our study period, falling 0.12 percent between 2013 and 2016.[8]

3. The Intervention

The Colombian government contracted the National Productivity Center (Centro Nacional de Productividad, CNP) to design and implement the program. CNP is a non-profit that originally was funded and supported by Japanese technical cooperation and has been the recipient of training and in-house technical assistance to develop capabilities in implementing managerial consulting services such as Lean, Six-Sigma, etc. During its 15 years of experience CNP has developed a model of operation that has allowed it to support more than 4,000 Colombian companies in different areas of management, innovation productivity and competitiveness. CNP used two types of consultants for the intervention. The first were lead consultants, who were long-term employees of CNP with more than 10 years' experience, and experience managing teams, often in multinational firms. They led area consultants, who had to have had at least 5 years' experience, and specialized in a particular focus area such as logistics or finance. The direct cost of implementation of this program was approximately US$2.4 million.

[7] The report notes a nominal growth rate of 11.2 percent, which we deflate using the Colombian inflation rate taken from the World Development Indicators.

[8] Export data and sales data from DANE and are for the CIIU sector 2930 "Manufacturing of parts, pieces, and accessories for automobiles and their motors".

3.1 Diagnostic phase

All firms, including the control, received a diagnostic as the first phase. This was implemented on a rolling basis between June and October 2013. The diagnostic was carried out by a team of 6 consultants, consisting of a lead consultant and five specialists, one for each area (Logistics, Human Resources, Finance, Marketing and Sales, and Production). The diagnostic began with an opening meeting with top and middle management, and then each area specialists would have five days of meetings with the responsible manager in the firm for their area to evaluate the 141 individual management practices that form Anexo K. This forms the baseline management practices measure. The consultants would also examine the firms key performance indicators for the last three years (to the extent records existed), and work with the leader to finish with a report (improvement plan) that analyzed managerial practices for each area, the key performance indicators for each area, and recommended practices to prioritize. This diagnostic phase lasted 2 full-time weeks and cost 8,426,550 COP (US$3,553) per firm.[9]

The diagnostic identified priority practices to be improved by management with the accompaniment of the consultants. These practices were intended to be ones which required minimal capital investment, and which could be implemented reasonably quickly and were expected to lead to relatively rapid improvements in the firm. While these priority practices were individualized by firm, some of the priority areas for improvement in each of the five areas were common to many firms. These include implementing master budgets across areas, improving systems for tracking costs, defining explicitly the strategic objectives of each position in the plant, implementing plans to improve the skills of people in management roles, lining up sales and marketing plans with business strategy, and analyzing machine downtime and quality problems daily across different supervisors.

3.2 Individual Consulting Treatment

Assignment to treatment took place after the diagnostic phase, in November 2013. Firms assigned to the individual consulting treatment group then received individual support for a period of 6

[9] We use the average exchange rate over 2014-15 of 2372 COP = 1 USD for all currency conversions in this paper. Cost numbers are implementation costs, and exclude initial costs of intervention design, and additional costs of data collection for the impact evaluation. To the extent this data collection process also helps firms improve management, it could be considered another part of the intervention, and averaged a further US$20,000 per firm (including the control group). Note that our costs are the costs to the government, and so do not include the opportunity cost of time to the firms participating, nor any minor travel costs incurred by them in travelling to meetings.

months, in the time window between March and November 2014. They were assigned a team of five consultants, one for each of the five processes (logistics, human resources, finance, marketing and sales, and production), along with a leader.

The intervention began with an opening meeting that brought together the leaders within the firm responsible for each of these five processes, along with the six consultants to define the different roles and responsibilities and set out a work plan. Then each of the five area consultants would visit the firms and provide 20 hours of training to the person in the firm in charge of their respective area. This would involve a theoretical part with the goal of familiarizing the firm's management with modern management concepts and methods, complemented with practical exercises to apply these concepts to their firm. This was then followed by individual consulting to help the firms implement the improvement plan developed during the diagnostic phase. Every area would be covered by different consultants and with different schedules but would typically involve weekly meetings of four hours per visit, spread over three to five months. Once per month, the team would meet with the whole firm's management to discuss improvements and re-define priorities and next actions. The total consultant time was 500 hours, consisting of 100 hours of providing training, and then approximately 100 4-hour sessions per firm of individual consulting. The cost of this individual intervention was US$28,950 per firm receiving treatment.

Based on our discussions with firms and own observations of the process, the implementation appears to have involved an emphasis on teaching firms how to measure and monitor key performance indicators, and on providing firms with the set of tools needed to better understand how their business is performing. It appears that there was less direct implementation from the consultants. For example, the consultants might go through the financial and performance data from the firm and suggest the need for the firm to consider new product lines or develop new markets abroad, but seldom make more direct recommendations (e.g. you should try exporting product X to Ecuador, or you should start using this production technology).

3.3 Group Consulting Treatment

The idea behind the group consulting treatment was to test whether the same gains in management improvements could be achieved more efficiently through working with small groups at a time, motivated in part through the way agricultural extension services are often implemented. The group treatment arm aimed to lower costs in two key ways. First, by working with multiple firms

at once, and potentially having them also learn from one another, each consultant's time was spread over more firms. Second, rather than consultants having to travel to the firms, most of the meetings took place in central meeting places such as conference rooms, cutting down on consultant travel time.[10] CNP had not previously done group consulting of this form, and designed the details of this intervention specifically for this program.

Groups were formed of 3 to 8 firms located in the same region, such that members are not direct competitors to one another, but are instead producing complementary products with similar management problems.[11] These groups were formed after the randomization, in November 2013. However, unfortunately a different government budgetary entity was designated to pay for this treatment arm than that was paying for the individual treatment. This entity significantly delayed the payment, meaning that the group intervention was unable to start until over a year after the individual intervention, running six months from September 2015 to May 2016 (with different groups starting and stopping at different times, and a break over the Christmas period).

Leaders from the firms in a group signed an agreement to work together and help each other improve. Like the individual treatment, the group treatment began with training classes that covered theoretical aspects of management. The difference is that these classes were delivered to the group in a classroom setting, instead of one-on-one in the firm. Each firm would send the staff in charge of a particular area or production process along to that training session. For example, when financial training was performed, firms would send the people responsible for the firm's financial components to the training. These sessions lasted for a total of 40 hours per group, including a session on the topic of cooperation among firms.

This was then followed by group consulting sessions, designed to help firms implement the management improvements. In any given week, a group would discuss two areas, having one or two meetings focusing on a single area (for a maximum of four meetings a week per group). Only management with responsibilities over the area being discussed would participate in the meetings. The same two areas would be covered at the same time over about eight weeks. After a break over

[10] This does shift some cost from the program to the firms. An approximate estimate is that the value of manager time spent travelling and the cost of manager transportation may have been up to $1,000-$2,000 per firm.

[11] Given the heterogeneity in products produced by the firms, in practice what mattered when forming groups was geography rather than competition – firms were grouped with other firms in the same city or general area of the city to make it easier for them to travel to meet one another in the group meetings. The composed groups are 1 group of 8 firms, 4 groups of 7 firms, 2 groups of 6 firms, 1 group of 4 firms, and 1 group of 3 firms.

Christmas, the remaining three areas would be covered the same way. The order in which areas were discussed was not the same for each group.

The group meetings would focus on the implementation of the actions agreed in the improvement plans of each company. Within each group, each firm had to work on the improvement of the topic that had been prioritized for a number of firms in the group, unless the firm excelled already in that topic. Therefore, each firm would still be focused on the issues that had been prioritized in the Improvement Plan but its Action Plan would be updated to include relevant issues taken from the other firms' Improvement Plans. If a firm already excelled in topics that were central in other firms' Improvement Plans, it would be used as an example and its experience would be discussed in detail.

In the individual intervention, consultants were at the firm for all visits, so could directly see implementation attempts and problems and adjust their recommendations accordingly. In contrast, during the group intervention, it was more difficult to directly verify changes being made in logistics and production. This was solved by requesting firms to provide evidence of what they had implemented in the form of bringing photos to the group meetings. These photos were not used as part of the measurement of practices, but more as an input into discussions of how to implement the practices. In addition, firms in the group treatment still had a monthly one-on-one visit, which took place at the plant, when a consultant would meet with senior management, and one hour at the end of each meeting was used to visit the plant and review improvements.

This process enabled the group intervention to be significantly cheaper than the individual intervention, with an average cost of US$10,500 per firm receiving treatment. Firms received 408 hours of consultant time each, consisting of 40 hours group training, and 92 4-hour group sessions.

4. Take-up, Data sources, and Attrition

4.1 Take-up

The take-up rate for the individual intervention was 86.8%, with all 46 of the 53 firms which started this intervention completing it. The longer delay until beginning the group intervention reduced the take-up rate for this intervention, with 40 of the 53 firms in this group (75.4%) starting the intervention, and 36 firms (67.9%) completing it. Table A5.1 shows the baseline characteristics of those who completed the intervention are not statistically different from those who dropped out,

with the one exception being that dropout from the individual treatment was more common in the Antioquia region than elsewhere. The main reasons given for drop-out from both groups were lack of owner time to participate, and lack of continuity in the program (especially for the group treatment).

4.2 Data Sources, Measurement of Key Outcomes, and Attrition

Baseline data were collected from the application form and diagnostic phase and cover firm characteristics in 2013. We then use three types of follow-up data, discussed in detail in Appendix 4. The first is data on the management practices in the firm. Our main measure is the Anexo K management score, which is a score measuring the average adoption rate of the 141 different practices detailed in Appendix 4. This was collected by CNP during in-person visits to the firms, with high scores of 4 or 5 double-checked. It was measured during the diagnostic for all 159 firms, monthly from the treatment groups during the time of their interventions, as well as annually in 2014 and 2015 for the individual and control groups, and in 2015 and 2016 for the group treatment. The second type of data consists of key performance indicators (KPIs) from the firms, which were collected during in-person visits. We hired Innovations for Poverty Action to provide an independent check and assistance in collecting data directly from the firms – this included oversight of both the management practice data and the KPI data. However, obtaining performance data directly from the firms was difficult and complicated by breaks in CNP's contracts and the long length of our project, which meant that many firms who had initially cooperated refused to provide data after several years. Appendix 5 discusses the attrition in these measures. We use this KPI data to measure impacts on defect rates, and as a supplement to our analysis of impacts on employment and firm sales.

Our third type of data, and main source for firm outcomes, comes from linking firms to administrative data sources. These data sources have the advantage of being collected over longer time periods, independently from the program, and with much less attrition than our attempts to collect performance data directly from firms. We used our partnership with the Colombian Planning Ministry (DNP) to link our firms to two administrative data sources. We use employment outcome data come from the PILA (*Planilla Integrada de Liquidación de Aportes* (Unified Register of Contributions)), which is the national information system used by firms to file the mandatory contributions to health, pensions, and disability insurance paid for workers. This data

has the advantage of covering almost all firms, since we could match 157 of the 159 firms to these records. Moreover, it is more comprehensive in length, enabling us to track firms at the monthly frequency from pre-intervention (January 2013), right through until the end of December 2018, which corresponds to three years after the group interventions and more than four years after the individual intervention ended. The potential drawback is that it only covers formal employment. Appendix 9 discusses this data in more detail, and compares it to the employment data we directly obtained from firms, finding a correlation of 0.93 (Figure A9.1), and that few firms appear to have large numbers of informal workers.

Secondly, staff from the Colombian statistics agency (DANE) linked the firms in our experiment to their database of the annual manufacturing survey (*Encuesta Anual Manufacturera* (EAM)) which is mandatory for establishments with more than 10 employees. They were able to locate 120 of our 159 experimental firms in this annual panel, with Appendix 3 showing that smaller and younger firms were less likely to be matched. These data provide annual sales, value-added, profits, and labor productivity measures from 2010 through to 2018, with a balanced panel of 100 firms appearing in all nine years.

5. Impact on Management Practices

The interventions aimed to improve specific management practices covered under the 141 practices that comprise Anexo K. These practices were measured for all firms during the diagnostic phase in 2013, and then measured monthly during the implementation periods of the individual and group interventions, and again one-year post-intervention. The control group had these measured towards the end of the individual treatment intervention, and again at the time of the one-year follow-up.

5.1 Overall Impact on Management

Figure 1 shows the trajectory of impacts on management practices for the overall Anexo K management score, and for the scores under the five separate areas of finances, human resources, logistics, marketing and sales, and production practices. The control group shows a gradual improvement in practices over time, which we attribute to the diagnostic and our data requests. We see that the individual treatment group sharply improves practices overall, and in all five areas, during the implementation phase, while the control group improves by much less. The group treatment likewise sharply improves practices for this treatment group during the implementation

phase, and ends up with practices at or above where the individual treatment group ended. This improvement in management then persists during the following year for both groups. Figure 2 compares the distributions of management practices at baseline, and at the last follow-up, for the three groups. Kolmogorov-Smirnov tests show we cannot reject equality of distributions at baseline, but at the endline, both the individual and group treatments are significantly different from the control group (p-values 0.004 and 0.003 respectively), although are not significantly different from each other (p-value 0.643).

For our regression analysis, we therefore classify our data into three periods: baseline, during the intervention (measured at the end of implementation for the individual and group treatments, and the first follow-up for the control group), and post-intervention (measured at the one-year follow-up post-intervention for the individual and group treatments, and the second follow-up for the control group). This time-shifts the data for the group treatment to account for the delay in implementation, which meant that its follow-ups took place a year later than the other two groups. We then estimate the following ANCOVA regression (McKenzie, 2012) for *t=2* (during) and *t=3* (post-intervention) that controls for the randomization triplets and the baseline level of management practices, and allows the impacts to vary during the intervention from post-intervention:

$$AnexoK_{i,t} = \alpha + \beta_1 Individual_i * During_t + \beta_2 Individual_i * Post_t$$

$$\gamma_1 Group_i * During_t + \beta\gamma_2 Group_i * Post_t + \sum_{g=1}^{53} \delta_g 1(i \in g) + \theta Post_t$$

$$+\rho AnexoK_{i,1} + \varepsilon_{i,t} \qquad (1)$$

Where *Individual* and *Group* denote assignment to the individual and group treatments respectively, *During* is a dummy for t=2, *Post* a dummy for *t=3*, $1(i \in g)$ is a dummy for firm *i* being in randomization triplet *g*, and the standard errors are clustered at the firm level.[12]

Table 2 presents the estimated treatment effects on these management practices. Panel A uses the unbalanced panel, which includes firms whose practices were measured in only one of the two follow-up periods, and Panel B the balanced panel of firms measured in both follow-ups. Four key results are evident. First, we see the immediate treatment impacts seen in Figure 1 are statistically

[12] Similar results for the impact on management practices are obtained if we instead use a firm fixed effects regression.

significant at the 1 percent levels for both treatments. Second, these treatments persist for at least one year post-intervention. The estimated effect size is between 8 and 10 percentage points, relative to the control group implementing 56 percent of the practices by 2015. Second, the impact persists. Third, the individual and group treatments yield impacts that are similar to one another in magnitude, and we cannot reject equality of treatment effects for the overall index, or for any of the five areas, in the post-intervention period. Appendix 6 shows these results are robust to using different weighting schemes to aggregate the individual practices into an index, and does not appear to be driven by sample attrition.[13]

How large an effect is this improvement of 8 to 10 percentage points in management practices? It is only approximately one-third the size of the improvement of 26 percentage points found by Bloom et al. (2013) from their management intervention in India[14], but approximately twice the size of the typical improvement found in standard business training courses given to smaller firms (McKenzie and Woodruff, 2014).

5.2 Which Practices Improved?

The improvement in management practices is broad, occurring in Figure 1 and Table 2 across all five areas with reasonably similar magnitudes. Table A7.1 looks at the sub-index and individual practice level. The individual treatment has a positive and statistically significant impact (at the 5% level) on 23 out of the 35 sub-indices (66%), and 65 out of the 141 individual practices (46%), while the group treatment has a positive and statistically significant impact (at the 5% level) on 20 out of the 35 sub-indices (57%), and 73 out of the 141 individual practices (52%). Table A7.2 examines which practices have had the largest impacts. These are mainly practices concerning defining strategic goals and objectives, setting up master budgets, and monitoring key performance

[13] Appendix 8 discusses our efforts to efforts to also measure changes in management using the World Management Survey (WMS) and Management and Organizational Practices Survey (MOPS). These measures are at a more general level than the Anexo K measures. A combination of budget constraints and attrition mean that we only have this data for 70 of the 159 firms (WMS), and 95 firms (MOPS). We show that our Anexo K measures are correlated with the WMS and MOPS in the cross-section, but not in the panel, and that our WMS and MOPS measures appear to be noisily measured, with less predictive power for business outcomes than Anexo K. Our measured treatment impacts on these two measures are smaller in magnitude and not statistically significant.

[14] In terms of standard deviations, our treatment impact is 0.8 to 0.9 s.d., compared to 1.58 s.d. in Bloom et al. (2013). However, differences in the heterogeneity of firms across studies mean one should be cautious in comparing effect sizes expressed in terms of standard deviations.

indicators. The smallest number of improvements are seen in human resource practices and logistics practices.

Figure 3 plots the estimated treatment effects practice by practice for the individual and group treatments. The correlation is 0.71, showing that the two different approaches to improving management not only resulted in a similar aggregate improvement in management, but also to a similar mix of practices improved. The main area of difference occurs with several production practices related to preventative maintenance, which improved more with the group treatment than the individual treatment.

Why didn't firms change more of their management practices? Qualitative interviews suggest several explanations. A first one is delays in implementation, which caused some firms to lose interest. The consultants pointed to problems getting family-run businesses to focus on improvements, and that a lack of a data culture prevents firms from recognizing their flaws. For this reason, much of their initial focus was on getting firms to collect KPIs and to have meetings to identify problems, which, in our opinion, may have come at the expense of "quick wins" in which changes in specific practices could be seen by firms to lead quickly to noticeable improvements in business outcomes.[15]

We also asked the consultants to go through a flowchart to explain why key practices identified in the diagnostic were not then implemented (before the intervention). This was done in early 2014 for approximately two practices per firm in 87 firms in the individual and control groups, for a total of 151 practices. Firms had heard of the practices, but were rated low in their knowledge about the practices, with 72% of firms being scored as a 1 or 2 out of 5 on knowledge of how to implement the practice. The consultants believed that external factors (<1%) and firm human and financial resources were not constraints to implementation (only 6%). In contrast, they thought that the firm owner mistakenly did not consider the practices to be profitable (net of any managerial or monetary costs of implementing) in 58% of cases. This is consistent with the findings of Bloom et al. (2013) that the main reasons for practices not being implemented were lack of knowledge about the practices, and firm owners not thinking the practices were worth implementing.

[15] For example, in India, the international consulting company we used started by identifying a couple of practices that could be changed quickly and where the firm could see immediate results, and then hand-held firms through changing these practices as a way to garner enthusiasm and momentum for broader changes.

5.3 Correlated Practice Changes Within the Group Treatment

The motivation for the group intervention suggested two possible ways in which working with firms in groups could foster improvements in management practices. A first possibility is one of coordinated experimentation and learning, whereby group members try to improve the same practice together, so are able to motivate and learn from one another. A second possibility is one of existing knowledge transfer, whereby group members are able to learn how to implement a practice from other group members who were already implementing it well to begin with. We explore the extent to which these two mechanisms are occurring in our sample by running the following regression for the change in management practice *j* in firm *i* assigned to group *g*:

$$\Delta Practice_{j,i,g} = \alpha + \beta \overline{\Delta Practice_{j,-i,g}} + \lambda \max_{-i,g} BaselinePractice_{j,-i,g} + \varepsilon_{j,i,g} \qquad (2)$$

Where $\overline{\Delta Practice_{j,-i,g}}$ denotes the mean change in practice *j* for other members in *i's* group, and $\max_{-i,g} BaselinePractice_{j,-i,g}$ denotes the maximum level of practice *j* at baseline among other members in *i*'s group. We stack the 141 individual practices, and then cluster the standard errors at the firm level. Note that since we have only a small number of groups, and which group from a firm ended up in was not randomly assigned, we view this analysis as descriptive and providing suggestive evidence to help explore mechanisms.

Table 3 reports the results of estimating equation (2). Column 1 shows that there is a significant positive association between the change in a practice for a firm and the mean change made by other firms in their group. Column 2 shows that, in contrast, there is no significant relationship with the highest baseline level of practices observed amongst other firms in the group. Columns 3 and 4 investigate whether it is the absolute, rather than relative, level of practices observed in another firm in the group that matters. Only 13 percent of firm*practice observations are for firms placed in a group with a high performer that already had this practice at the maximum level of 5. Column 3 finds a positive, but statistically insignificant association between having such a high performer and changing the practice. Column 4 similarly looks at whether there is a firm that is already doing the practice at the level of 4 or 5, and again finds no significant association. Columns 5 and 6 control for both factors together, and confirms the significant and positive association with the average change made by others in the group. A one-unit change (on a 5-point scale) in the practice by others in the group is associated by a 0.1unit change by the firm. This suggests some

coordinated experimentation and learning is taking place within groups, but that group members are not taking existing best practices from other group members across into their own firms.

6. Impacts on Firm Outcomes

6.1 Impact on Employment

Employment is a key outcome for several reasons. First, from the policy side, governments around the world are interested in increasing employment in larger, more formal firms. This is the case in Colombia, where the unemployment rate averaged 8 percent during our intervention period, and where 47 percent of those who were employed were in informal jobs. As shown in Appendix 9, most of the employment in our firms is formal and eligible for social security and health benefits, and the mean (median) monthly wages of firms in our sample of $492 ($331) in 2018 are well above the minimum monthly salary of $248 and median monthly wage of $283.[16] Second, from a measurement perspective, employment is a key measure of firm growth. This is both a result of data coverage (monthly formal employment data are available for more firms and over a longer time period than any of the survey outcomes we consider), and of the inherent volatility in firm sales (Lewis and Rao, 2015). For these reasons, employment is also the main measure of firm growth that Bruhn et al. (2018) highlight in their individualized consulting experiment. Finally, from a theory perspective, employment growth is a key marker of firm size and productivity as firms age (e.g. Hsieh and Klenow, 2014).

Given the heterogeneity amongst firms in initial employment size, and the differences in coverage of the different data sources, we use firm fixed effects in estimating the treatment impacts. We estimate the following equation for firm *i* at time *t:*

$$Employment_{i,t} = \alpha_i + \beta_1 Individual_i * During_{i,t} + \beta_2 Individual_i * Post_{i,t}$$

$$\gamma_1 Group_i * During_{i,t} + \gamma_2 Group_i * Post_{i,t} + \sum_{s=1}^{T} \delta_s 1(s = t) + \varepsilon_{i,t} \qquad (3)$$

Where the α_i are firm fixed effects, *During* and *Post* indicate the periods during the individual or group interventions, and after these interventions respectively, 1(s=t) are time fixed effects, *Individual* and *Group* denote assignment to the individual and group treatment status respectively, and the standard errors $\varepsilon_{i,t}$ are clustered at the firm level. The randomization triplets are subsumed

[16] 2018 numbers use an exchange rate of 3155 COP to 1 USD. The minimum monthly salary in Colombia for 2018 was 781,242 COP, and median monthly wage was 882,500.

by the firm fixed effects here. We consider both levels and the inverse hyperbolic sine of employment as outcomes. β_2 and γ_2 then give the average impact of the individual and group treatments respectively over all available post-treatment periods.[17]

Table 4 presents the treatment impacts on employment. The first two columns use the employment data obtained from firms. While these data are available for some 145 firms for some months, only 108 of the firms have data for much of 2017. The group treatment results in a statistically significant increase in employment of 6 workers post-treatment, or 12 percent. In contrast, the individual treatment results in negative point estimates on the level of employment, and an effect which is significantly different from the group treatment at the 10 percent significance level when employment is measured in levels, but not significantly different for the I.H.S. transformation.

Columns 3 to 6 of Table 4 use formal employment data from the PILA. There is no significant impact of treatment on firm survival (Appendix 5). Columns 3 and 4 show that when we consider the employment levels of surviving firms, the group treatment firms are significantly larger in size, with similar magnitudes as found with the firm data. In contrast, the individual treatment has smaller impacts on employment, which are not significantly different from zero post-treatment, and which are significantly different from the group treatment when using the inverse hyperbolic sine. Columns 5 and 6 consider unconditional employment as the outcome, coding employment as zero once firms die. The point estimates still suggest a 4 worker (16 percent) increase in employment after the group treatment, but the standard errors are larger, and these impacts are no longer statistically significantly different from zero, or from the individual treatment.

Columns 7 to 10 of Table 4 report impacts on total employment from the EAM. Columns 7 and 8 use the unbalanced panel of all 120 firms that are ever found in the EAM, while columns 9 and 10 use the balanced panel of 100 firms that have data in all nine years from 2010 to 2018. In both cases we find a significant impact of the group treatment on employment post-treatment. The point estimates are larger than with the PILA, indicating a 13 to 15 worker increase, or 25 to 30 percent in total employment. This difference in magnitude reflects both the EAM not containing all of our smaller firms, for which the point estimates are smaller, and that the EAM also includes temporary and contract workers that are not directly linked to the firm in the PILA. The impacts of the

[17] Appendix Table A10.1 presents impacts year-by-year and shows that we cannot reject equality of impacts across post-treatment years.

individual intervention are approximately half the magnitude of those of the group intervention, at 7 workers or 8-12 percent. These impacts are not significantly different from zero, but neither can we reject equality of the individual and group impacts in the EAM data.

An interesting question is where the increased employment in the group treated firms comes from and whether it changes the composition of the labor force in these firms. We were able to use anonymized worker-level data from the PILA to examine these questions in more detail for formal workers (see also Appendix 9). A first point to note is that there is considerable worker churn: there are 23,156 distinct workers who work at least one month in one of our firms in the 2013 to 2017 period, but only 7,500 to 8,000 workers in any given month. On average firms have 3 percent of their workforce leave each month and 3 percent join. Most of this churn comes from outside of the study firms: only 272 workers (1.2%) worked for two or more firms in our sample during this five-year period, and only 32 workers worked for firms in more than one treatment group. The growth in the group treatment firms therefore did not come from them hiring away workers already working in the other treatment groups. Table A9.1 then examines the impact of treatment on the composition of workers, finding no changes in the gender or age of workers with treatment, no significant change in worker compensation, and that the group treatment firms retained more of their workers (significant at the 10 percent level). Table A9.2 shows that the few workers who switch between study firms are similar in age and salary to those who do not switch, suggesting the worker flow is not disproportionately made up of more technical or managerial workers.

6.2 Impact on Firm Performance

Linking the firms to the EAM enables us to use nine years of annual data from 2010 through 2018 in estimating firm outcomes. We estimate equation (3), which controls for firm and time fixed effects, and show results for both the full set of 120 firms that were able to be linked to the EAM, as well as for the balanced panel of 100 firms present in all nine years of the data. We examine four measures of firm performance: annual sales (columns 1 and 2), annual profits (columns 3 and 4), value-added (columns 5 and 6), and annual production (columns 7 and 8). To account for multiple testing, we also construct an aggregate index of standardized z-scores of these different outcomes and report impacts on this performance index in columns 9 and 10.

We see that the group intervention had positive and statistically significant impacts both during and post-intervention on all four measures of firm performance, and for the overall index measure.

The results are similar for the unbalanced and balanced panels, and are significant in both levels and logs (with the exception of profits). The point estimates imply a 28 to 33 percent increase in sales and production in the post-intervention period, a 43 percent increase in value-added, and a 5 to 26 percent increase in profits. In level terms, the annual increase in sales is 1,705 million pesos (USD 720,000) and the level increase in profits is 647 million pesos (USD 273,000), with a 95 percent confidence interval for the increase in profits of USD17,400 to USD528,000. This is the average effect across all three years of post-intervention data. Table A10.1 shows we cannot reject equality of impacts by year after intervention.

The impacts of the individual treatment on firm performance are almost all positive in the post-treatment period, but not statistically significant for any of the four outcomes. The magnitudes of the estimates are smaller than those of the group treatment, and more sensitive to the choices of levels versus logs and unbalanced versus balanced samples. For the levels specification, we cannot reject that the impact of the individual treatment on the overall performance index measure is the same as that of the group treatment, but neither can we reject the null of no treatment effect. The impact of the individual treatment on the overall index measure is statistically significant at the 5 percent level for the inverse hyperbolic sine/logs specification, but only for the balanced panel. Taking the balanced panel results, 95 percent confidence intervals for the impact on sales are (in levels) [-1166 million pesos, +2801 million pesos]; (in logs): [-7%, +38%]. The increase in profits has a 95 percent confidence interval of -27 percent, + 27 percent.

6.3 Potential Channels of Production Impact

The results on employment and firm performance show that the group intervention increased the size of the firm, causing it to employ more people, sell more, and earn higher profits. The individual intervention had positive point estimates, that are not significant and less conclusive. In Table 6 we examine several potential mechanisms through which these changes in outcomes may have occurred. Column 1 considers the defect rate. Bloom et al. (2013) found quality improvements to be one of the first signs of improvement from better management in their Indian study. We only have defect data in 2017 for 78 of the firms in the study, due to many firms not keeping consistent records on defects. A first point to note is that the defect rates are low (which is one reason some firms do not record them). The control group has a mean defect rate of 0.025 and median rate of 0.007 in 2017, which compares to much higher defect rates in India (5 percent

of output was scrapped, *after* mending of defects was done). The result is that many of the auto parts firms do not have much scope to reduce defects, and we see treatment effects that are all very close to zero and statistically insignificant.

The remaining columns use data from the EAM.[18] Column 2 examines the treatment impact on inventory levels. In India, Bloom et al. (2013) found firms had excess inventory levels, which they reduced when management improved. Large stockpiles of inventories are less common in the auto parts sector, with some firms doing job work and producing upon request. The control mean level of inventories is equal in value to just over two months of mean sales. We see that no significant impact of either treatment on inventories, although the confidence intervals are wide.

Bloom et al. (2010) find that better management is associated with firms using less energy use, and suggest that this could arise from less wastage occurring through lean manufacturing, as well as through implementation of energy-saving ideas. The interventions did not stress energy efficiency, so any effect would likely be through a general effect on less wastage rather than through energy-saving techniques specifically introduced. Column 3 examines whether this mechanism is occurring in our experiment, looking at the ratio of energy costs to sales. Total energy expenses increase as the firms produce more, and this ratio does not significantly change.

We then turn to labor productivity, measured as value-added per worker. Since this is a key outcome, we report results for both the unbalanced and balanced panels in columns 4 and 5. From Tables 4 and 5, we see that the group intervention increased both value-added and employment, with the magnitude of the increase in value-added higher than that of employment. The consequence is that the treatment effect on log value-added per worker is 11 to 14 percentage points, although this is not statistically significant (95 percent confidence interval [-15%, +45%] for the balanced panel). These point estimates are consistent with the possibility that the firms grew in part by getting higher productivity from each worker, as well as from using more workers. The impact on labor productivity from the individual treatment is more sensitive to the choice of unbalanced versus balanced panel, ranging from 4 to 10 percent, although in neither case can we reject equality with the group treatment in the post-treatment period. A 95 percent confidence interval for the balanced panel impact post-treatment is [-15%, +43%].

[18] Given that the results are similar for the unbalanced and balanced EAM panels, we present results for the unbalanced panel which uses all available information.

6.4 Cost-Benefit and Comparison to Policy Maker Expectations

Both the individual and group treatments succeeded to a similar magnitude in improving the set of management practices measured by the Anexo K. The impacts on firm performance outcomes show increases in firm size for the group treatment that are larger in magnitude than those of the individual treatment effects, although only in some cases is this difference statistically different. The group treatment cost USD10,500 per firm for the intervention stage, compared to USD28,950 per firm for the individual treatment. The group treatment therefore clearly dominates the individual treatment on a cost-benefit basis from the point of view of government finances. Even if we include an additional USD1,000-2,000 in travel costs that firms in the group treatment may face compared to those in the individual treatment, the group treatment still dominates.

Using the EAM data we estimate a mean increase in annual profits from the group treatment of USD273,000, with the 95 percent confidence interval ranging from USD17,400 to 528,000. The intervention would then pay for itself within one month of returns at the point estimate, and within one year at the bottom of this confidence interval. These cost-benefit calculations would look less promising from a government policy perspective if the gains to treated firms came from them capturing sales from control firms or from other firms outside of the experimental sample. At least within our experimental sample, firms specialize in different products (which is what allowed groups to be formed easily without having firms who are competitors), suggesting that internal validity of our estimates should not be invalidated by such spillovers. Moreover, as noted in our discussion of the setting, the sector is one where the main competitors to most firms are imports, which became more expensive with the depreciation of the peso. It therefore seems likely that any sales gains achieved by the group treatment would have mostly come from taking business away from imports.

In June 2014, we elicited expectations about the program's impact on employment and productivity from 15 policymakers drawn from the Ministry of Planning (DNP), Ministry of Commerce and Tourism, SENA, and Program of Productive Transformation (PTP). The expected mean (median) treatment effect for the individual treatment was 5.7% (3%) for employment and 16.3% (10%) for productivity; while for the group treatment the expected mean (median) treatment effect was 3.3% (5%) for employment, and 7.3% (5%) for productivity. We also asked what size impacts they would require to consider the program a success that could be scaled at the national level: the mean response was 6% for employment for both programs, and 24% for the individual

program on productivity, and 13% for the group program. Our estimated impacts of the group treatment on employment exceed these policymaker expectations and the benchmark they had set for scale-up, whereas for productivity the desired impact lies just above our point estimate, but well within the confidence interval.

6.5 Why did the group treatment do better than the individual?

The group and individual treatments led to similar improvements in management practices, yet we find stronger evidence of improvements in firm outcomes for the group treatment. What explains this difference? A first possibility is that the two treatments did have similar effects, and it is just small sample sizes coupled with firm heterogeneity that prevents us from detecting this effect in the individual treatment group. Looking just at the level impacts is supportive of this viewpoint, since we find positive and significant impacts of the individual treatment on all of our firm performance measures in the EAM, and cannot reject equality of the effect on our aggregate performance index with that of the group treatment. However, when we look at outcomes in logs, then we can reject equality of the individual and group treatments for several firm performance outcomes and our overall index, especially when using the unbalanced panel. This suggests that there may be a small number of individual treatment firms that benefited a lot, pulling up the level effects, but that the average firm may have benefited less from being in the individual treatment than the group treatment.

A second possibility is then that the group treatment actually did have a larger impact. This could be because it either provides a way for the improvements in management to persist longer (beyond the period we measure management practices over), or because it delivers additional benefits to firms beyond the improvements in measured management practices. While the Anexo K is fairly comprehensive in measuring a wide set of management practices, they may not capture all changes occurring. To investigate this possibility, group firms were asked approximately one year after the intervention whether they still met with other group members, and what the main benefit of meeting in a group had been. None of the firms continued formally meeting together as a group, but 54 percent said they still communicate occasionally with other group members. The main benefit they saw of meeting in a group was to interchange experiences, noting the value of seeing other firms facing similar problems, and how others had solved these problems. Only four firms said they saw a possibility of using the group to find a supplier or customer, with only one giving an example of this actually happening, saying it was short-lived. This suggests that if the group

treatment is having an additional effect, it is more through providing advice and specific solutions to problems firms face (as in Brooks et al, 2018) or experiencing directly how others implement better managerial practices, and reducing uncertainty about their usefulness, rather than through direct business relationships.

Fostering linkages between the firms may be particularly valuable because there were very linkages across firms before our intervention, and limited knowledge sharing with other firms. A 2013 survey of 138 of the 159 firms in our experiment by *Infometrika* asked firms to list all other firms that they interact with about business. Out of the 21,804 possible bilateral linkages (138 firm who could each have listed any of the other 158 firms), we find only 15 linkages (0.07%). This makes spillovers across treatment groups unlikely (only one control firm knew any firm in the group treatment, and only three control firms knew firms in the individual treatments). When asked if they shared information or knowledge with any other firm, the majority of firms pre-treatment did not share information on production practices (70%), personnel practices (80%), logistics (82%) or quality control (74%). In this low information-sharing environment, the linkages formed by the group treatment may have been valuable beyond their effect on the management practices that we could measure.

Our sample size and available data on other channels make it hard for us to completely judge between these two different possibilities. Indeed, both may hold – there may be no average difference in the impacts of the two treatments on firm performance after similar improvements in measured management, but the group intervention may have still benefited some firms in ways the management practices do not capture. Given the cost advantage of the group treatment, we believe it deserves further replication and experimentation, with larger samples and innovations in measurement hopefully enabling closer examination of other potential channels of impact.

7. Conclusions

The experiment of Bloom et al. (2013) provided a proof-of-concept that poor management could be improved. But moving from a pilot demonstration to a scalable program of management improvement requires lowering the cost of delivery and testing whether such a program can be locally implemented when subject to the constraints imposed by government bureaucracy. As is common with other social programs (Rossi 1987, Vivalt 2019), impacts on management are smaller when delivered by program run by a government at scale than under a small researcher

pilot. Yet, both the individual and group treatments were able to improve management practices by 8 to 10 percentage points, with this resulting in an increase in firm performance and firm productivity. The group treatment model pioneered here clearly dominates the individual consulting model on a cost-benefit basis and offers a promising approach to scaling management. Part of its success, independent of cost, appears to arise precisely from the kind of group learning effects highlighted in recent literature that may more than offset the reduced individual attention by professional consultants.

While we are able to demonstrate that the group consulting treatment works, our sample size and number of groups prevents us from being able to offer detailed recommendations as to for what types of firms it best works, or for how groups should be optimally formed. An important area for future work will be to test this idea with many more firms and groups, and randomly vary the composition of groups as was done by Cai and Szeidl (2018). Future work can then also attempt to further measure other channels through which the group intervention improves firm performance.

As with firms, good management also matters for the public sector (Rasul and Rogger, 2018), and there were several challenges to implementation. These included delays in contracts which caused challenges for data collection, and delays in implementation which likely reduced the effectiveness of the programs implemented. It is also possible that contracting only a single organization to implement the intervention may have led to hold-up problems and removed the performance incentives that competition among consulting firms could have provided. A Government contemplating scaling up management support programs in the least costly way therefore should consider the group extension approach, and pay careful attention to the quality of their own management in doing so.

References

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukerji, Marc Shotland and Michael Walton (2017) “From Proof of Concept to Scaleable Policies: Challenges and Solutions, with an Application”, *Journal of Economic Perspectives* 31(4): 73-102.

BBVA Research (2018) “Situación Automotriz 2018 Colombia”, BBVA Research, March.

Bloom, Nicholas, and John Van Reenen (2007). "Measuring and Explaining Management Practices across Firms and Countries" *Quarterly Journal of Economics, 122*(4), 1341-1408.

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, Aprajit Mahajan, David McKenzie, and John Roberts (2020) "Do Management Interventions Last? Evidence from India", *American Economic Journal: Applied Economics*, 12(2): 198-219,

Bloom, Nicholas, Christos Genakos, Ralf Martin and Raffaella Sadun (2010) "Modern management: good for the environment or just hot air?", *Economic Journal* 120(544): 551-572.

Bloom, Nicholas, Raffaella Sadun, and John Van Reenen (2016) "Management as a Technology", NBER Working Papers 22327, National Bureau of Economic Research, Inc..

Bloom, Nicholas, Erik Brynjolfsson, Lucia Foster, Ron Jarmin, Megha Patnaik, Itay Saporta-Eksten, and John Van Reenen (2019) "What Drives Differences in Management Practices?", *American Economic Review*, 109(5): 1648-83.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a and Justin Sandefur (2018) "Experimental Evidence on Scaling Up Education Reforms in Kenya", *Journal of Public Economics*, 168: 1-20.

Brooks, Wyatt, Kevin Donovan and Terence Johnson (2018) "Mentors or Teachers? Microenterprise Training in Kenya", *American Economic Journal: Applied Economics*, 10(4): 196-221.

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): 635-87.

Cai, Jing and Adam Szeidl (2018) "Interfirm Relationships and Firm Performance", *Quarterly Journal of Economics* 133(3): 1229-1282.

Chandler, Alfred (1977) *The Visible Hand: The Managerial Revolution in American Business*. Harvard University Press: Cambridge, MA.

Chatterji, Aaron, Solene Delecourt, Sharique Hasan and Rembrand Koning (2018) "When does Advice Impact Startup Performance?", *Strategic Management Journal*, 40(3), 331-356

Dalton, Patricio, Julius Rüschenpöhler, Burak Uras and Bilal Zia (2018) "Learning Business Practices from Peers: Experimental Evidence from Small-Scale Retailers in an Emerging Market", https://pure.uvt.nl/portal/files/23354244/2_WP_Dalton_et_al_Learning_from_Peers_DFID.pdf

Giorcelli, Michela (2019) "The long-term effects of management and technology transfers", *American Economic Review* 109(1): 121-52.

Higuchi, Yuki, Vu Hoang Nam and Tetsushi Sonobe (2017) "Management skill, entrepreneurial motivation, and enterprise survival: Evidence from randomized experiments and repeated surveys in Vietnam", Mimeo. https://www.canr.msu.edu/afre/uploads/files/Higuchi_Paper_1217.pdf

Hsieh, Chang-Tai and Peter Klenow (2014) "The Life Cycle of Plants in India and Mexico", *Quarterly Journal of Economics* 129(3): 1035-1084.

Lafortune, Jeanne, Julio Riutort and José Tessada (2018) "Role models or individual consulting: The impact of personalizing micro-entrepreneurship training", *American Economic Journal: Applied Economics*, 10(4): 222-45.

Lewis, Randall and Justin Rao (2015) "The unfavorable economics of measuring the returns to advertising", *Quarterly Journal of Economics* 130(4): 1941-73.

Londoño, Andrés (2017) "Low Productivity: the Elephant in the Room in Colombia's Minimum Wage Debate", *Panam Post*, November 28 https://panampost.com/andres-londono/2017/11/28/low-productivity-minimum-wage-debate/

McKenzie, David (2020) "Small Business Training to Improve Management Practices in Developing Countries: Re-assessing the evidence for "training doesn't work"", Mimeo. World Bank.

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 (2017) "Business Practices in Small Firms in Developing Countries", *Management Science*, 63(9): 2967-81

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

Proexport Colombia (2012) "Automotive Industry in Colombia", http://www.investincolombia.com.co/attachments/Automotive%20Industry%20in%20Colombia%20-%20April%202012.pdf [accessed February 16, 2015]

Rasul, Imran and Daniel Rogger (2018) "Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service, *Economic Journal* 128 (608): 413-446

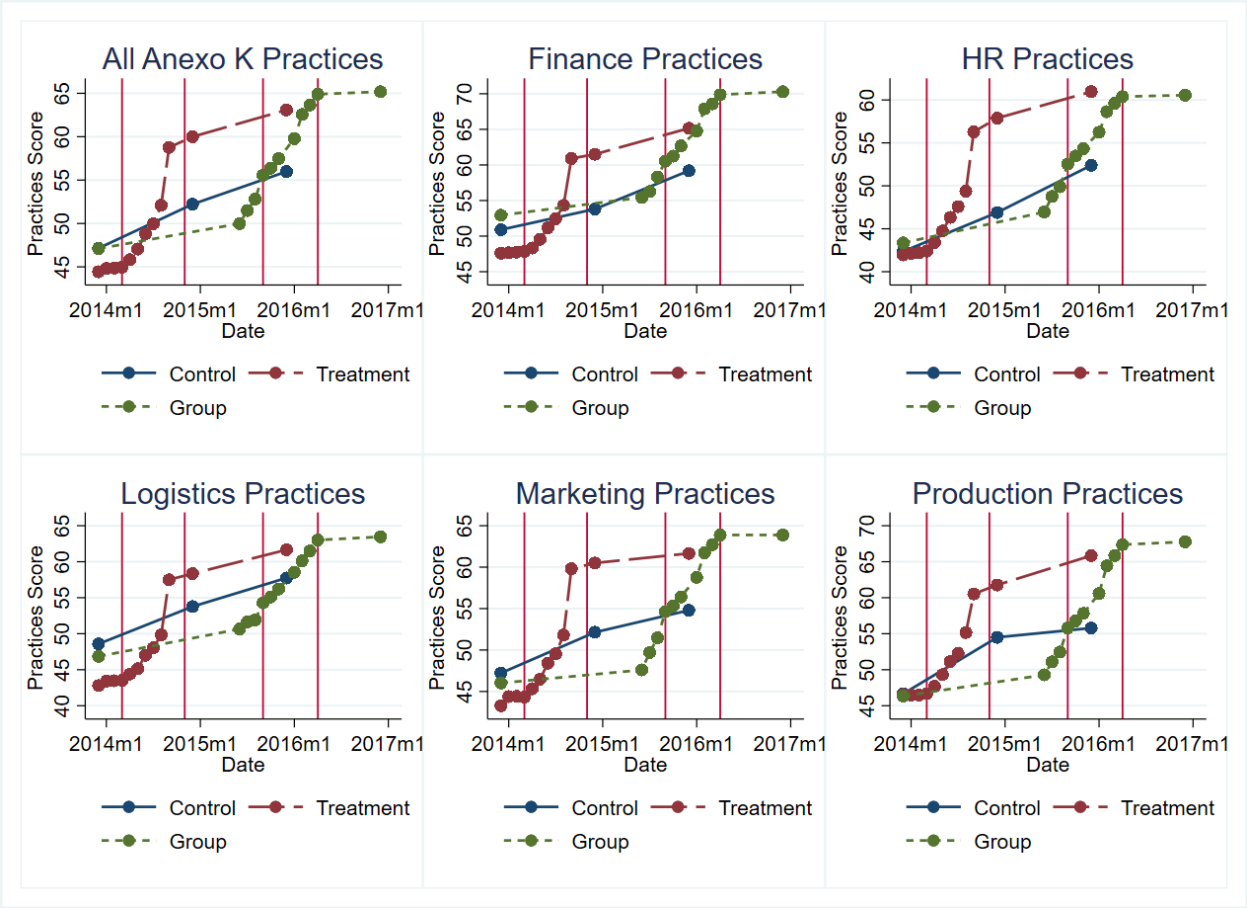
Ray, Debraj (2006) "Aspirations, Poverty, and Economic Change", pp. 409-422 in Abhijit Banerjee, Roland Benabou, and Dilip Mookherjee (ed.), *Understanding Poverty*. Oxford University Press, Oxford, UK.

Reina, Mauricio, Sandra Oviedo and Jonathan Moreno (2014) "Importancia Económica del Sector Automotor en Colombia", Fedesarrollo, Bogota.

Rossi, Peter (1987) ""The Iron Law Of Evaluation And Other Metallic Rules", pp. 3-20 in Joan Miller and Michael Lewis (ed.) *Research in Social Problems and Public Policy volume 4.* Jai Press Inc.

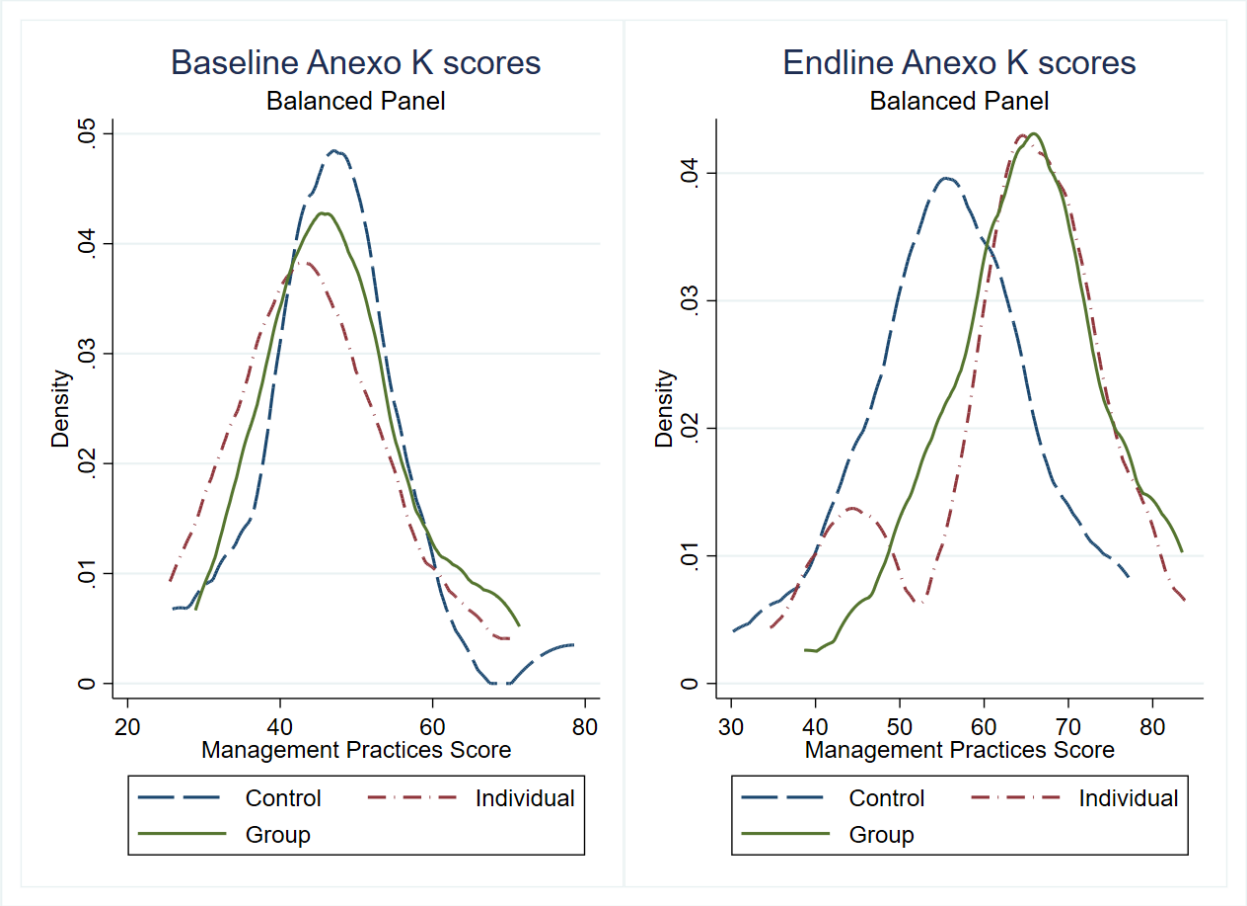
Vivalt, Eva (2019) "How much can we generalize from impact evaluations?", *Journal of the European Economic Association*, forthcoming.

Figure 1: Trajectory of Impacts on Management Practices



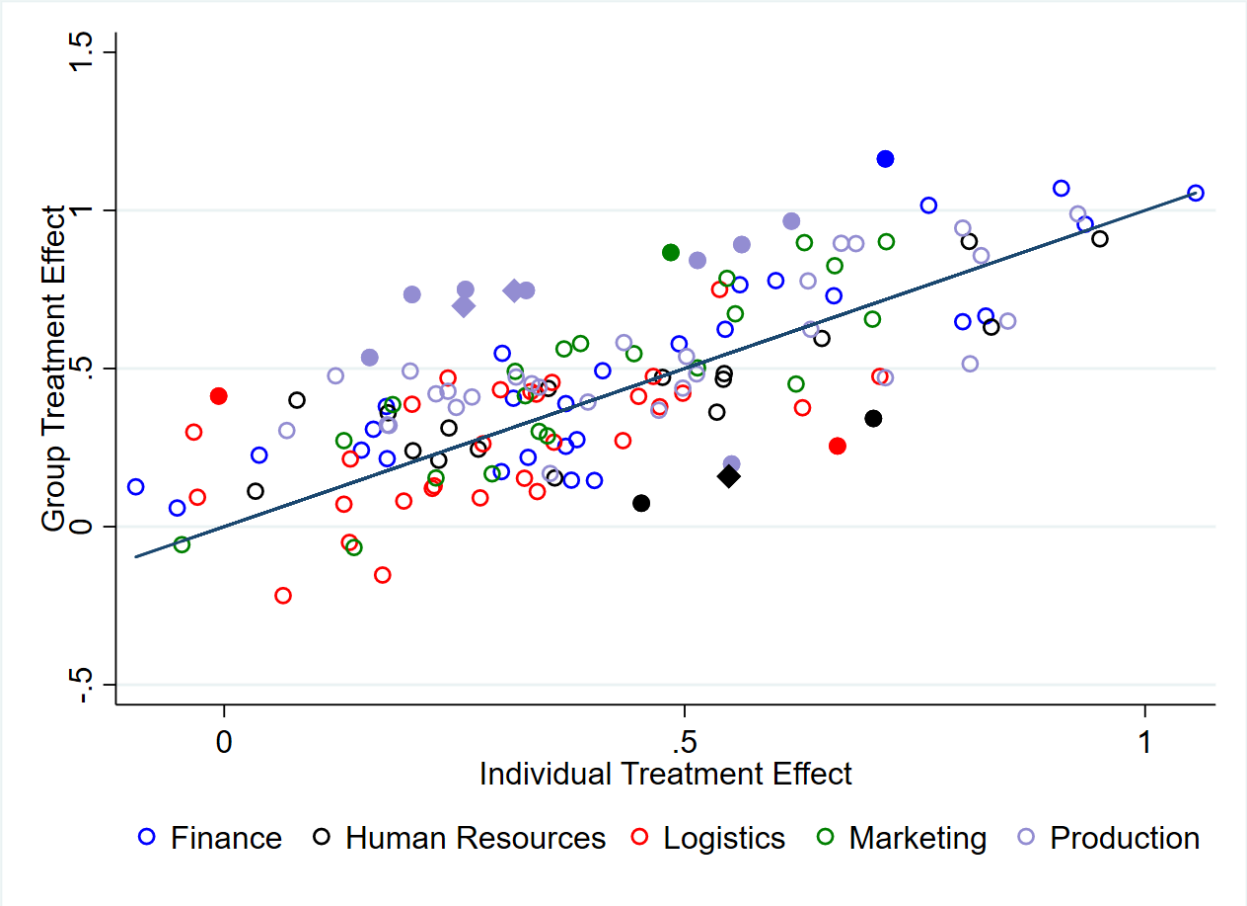
Notes: Means shown by treatment status. Anexo K was measured at baseline (2013) for all firms. It was then measured monthly during implementation of the individual and group treatments, along with a one-year follow-up, and was measured for the control group at the same time as the end of the individual intervention, and at the time of the individual one-year follow-up. Vertical lines indicate approximate periods of implementation of the individual intervention (first two lines) and group intervention (second two lines). Data are for the unbalanced panel, although figure looks similar for balanced panel.

Figure 2: Impact on Distribution of Management Practices



Notes: Kernel densities shown of Anexo K management practices at baseline, and at last follow-up, for the balanced panel of firms for which these practices were measured at all points in time. Kolmogorov-Smirnov tests of equality of distributions at baseline have p-values 0.210 (control vs individual), 0.998 (control vs group), and 0.422 (individual vs group); and at endline have p-values 0.004 (control vs individual), 0.003 (control vs group), and 0.643 (individual vs group).

Figure 3: The Individual and Group Treatments Improved Specific Practices to a Similar Extent



Notes: Empty circles denotes that difference between the two treatments is not statistically significant at the 5% level; Solid circles indicate that difference between the two treatments is statistically significant at the 5% level; Solid diamonds indicate that difference is statistically significant at the 1% level. Correlation between group treatment effect and individual treatment effect is 0.71. 45 degree line shown.

Table 1: Baseline Balance

	Overall Sample		Means by Treatment Group			p-value for testing equality		
	Mean	S.D.	Control Group	Individual Consulting	Group Consulting	Control v Individual	Control v Group	All 3 Equal
Variables used for matched triplets								
Number of Employees	59	53	64	61	53	0.726	0.106	0.219
Small Firm (<=50 employees)	0.59	0.49	0.60	0.58	0.58	0.276	0.276	0.516
Medium Firm (>50 employees)	0.41	0.49	0.40	0.42	0.42	0.276	0.276	0.516
Cundinamarca	0.48	0.50	0.55	0.49	0.40	0.526	0.085	0.212
Valle	0.16	0.37	0.17	0.09	0.23	0.234	0.421	0.132
Labor Productivity	31	18	26	32	34	0.059	0.017	0.042
Financing Practices	50	14	51	48	53	0.165	0.39	0.068
Human Resources Practices	43	12	42	42	43	0.854	0.634	0.780
Logistics Practices	46	13	49	43	47	0.016	0.441	0.052
Marketing Practices	46	15	47	43	46	0.111	0.63	0.245
Production Practices	47	13	47	47	46	0.962	0.882	0.989
Variables not explicitly balanced on								
Level 2 Supplier	0.94	0.24	0.94	0.94	0.92	1.000	0.705	0.911
Metal Products	0.60	0.49	0.75	0.51	0.53	0.005	0.009	0.007
Plastic Products	0.18	0.38	0.15	0.17	0.21	0.779	0.435	0.734
Firm Age (Years)	24	14	27	23	22	0.191	0.107	0.226
Anexo K score	46	10	47	44	47	0.124	0.978	0.219
USD Sales in 2013	2311669	2882942	1816503	2847673	2301337	0.046	0.211	0.084
Export at all in 2013	0.45	0.50	0.47	0.42	0.45	0.546	0.830	0.831
Sample Size	159		53	53	53			
Omnibus test p-value						0.151	0.636	

Notes:

P-values for testing equality of means for a variable y come from testing b1=0, b2=0, and b1=b2=0 in regression y=a+b1*individual treat+ b2*group treat + c'randomization triplet dummies + e

Omnibus test p-value comes from F-test of joint orthogonality of all variables X in regression individual treat= a + b'X + c'randomization triplet dummies + e

Table 2: Impact on Management Practices

	Overall Score	Finance Practices	HR Practices	Logistics Practices	Marketing Practices	Production Practices
Panel A: Unbalanced Panel						
Individual Treatment*During Intervention	9.703***	9.644***	10.793***	8.708***	10.637***	5.696***
	(1.370)	(1.852)	(1.822)	(1.603)	(2.280)	(1.806)
Individual Treatment*Post Intervention	9.620***	9.712***	8.974***	8.585***	9.451***	8.488***
	(1.830)	(2.413)	(2.508)	(2.457)	(2.466)	(1.993)
Group Treatment*During Intervention	11.971***	13.841***	12.249***	9.327***	11.899***	11.798***
	(1.660)	(2.057)	(2.078)	(2.047)	(2.599)	(1.993)
Group Treatment*Post Intervention	8.544***	9.820***	7.156***	5.860**	9.046***	10.694***
	(1.894)	(2.306)	(2.655)	(2.539)	(2.637)	(2.048)
Sample Size	225	226	226	225	226	225
P-value: Individual=Group During	0.145	0.027	0.451	0.753	0.568	0.002
P-value: Individual=Group Post	0.533	0.958	0.365	0.235	0.864	0.315
Control Mean	55.98	59.18	52.39	57.75	54.80	55.79
Control SD	10.79	13.79	11.25	14.33	12.58	11.19
Panel B: Balanced Panel						
Individual Treatment*During Intervention	9.861***	10.608***	11.111***	8.639***	9.072***	6.803***
	(1.756)	(2.277)	(2.328)	(1.962)	(2.985)	(2.010)
Individual Treatment*Post Intervention	9.757***	10.118***	9.463***	8.629***	8.568***	8.935***
	(2.014)	(2.650)	(2.780)	(2.646)	(2.723)	(2.078)
Group Treatment*During Intervention	12.118***	15.094***	12.227***	8.942***	11.309***	12.688***
	(2.029)	(2.373)	(2.583)	(2.413)	(3.349)	(2.279)
Group Treatment*Post Intervention	8.889***	9.912***	7.502**	6.022**	9.166***	11.513***
	(2.067)	(2.490)	(2.912)	(2.729)	(2.920)	(2.157)
Sample Size	202	202	202	202	202	202
P-value: Individual=Group During	0.152	0.027	0.555	0.881	0.341	0.006
P-value: Individual=Group Post	0.627	0.925	0.343	0.274	0.813	0.248
Control Mean	55.98	59.18	52.39	57.75	54.80	55.79
Control SD	10.79	13.79	11.25	14.33	12.58	11.19

Notes:

Panel A is for the 124 firms for which Anexo K management practices are measured post-baseline, panel B for the 101 firms for which practices are measured both during and after intervention.

Robust standard errors in parentheses, clustered at the firm level. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

Anexo K management practices are 141 management practices divided into five sub-areas.

Ancova estimation controls for baseline (December 2013) mean, and time fixed effects included, along with randomization triplet dummies.

Note: Group treatment moved back one period, since no control group data collected during 2016.

Table 3: Correlation of Practice Changes Within Groups

Dependent Variable: Change in Practice between Baseline and Endline

	(1)	(2)	(3)	(4)	(5)	(6)
Mean Change in Practice for other Group Members	0.100*				0.104**	0.102**
	(0.050)				(0.049)	(0.049)
Maximum Baseline Level of Practice for Other Group Members		0.001			0.014	
		(0.021)			(0.019)	
At least one other member has Practice at level 5 at Baseline			0.028			
			(0.059)			
At least one other member has Practice at level 4 or 5 at Baseline				-0.002		0.012
				(0.033)		(0.030)
Sample Size (Firms*Practices)	5069	5210	5210	5210	5069	5069
Mean Change in Practices	0.168	0.171	0.171	0.171	0.168	0.168

Notes:

Regression uses the stacked panel of 141 practices for firms in the group treatment.

Robust standard errors in parentheses, clustered at the firm level. *, **, and *** denote significance at the 10, 5 and 1 percent levels.

Table 4: Impact on Employment

	Firm Survey Data Jan 2013-Dec 2017		PILA Data Jan 2012-Dec 2018				EAM Data Annual 2010-2018			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
			Conditional		Unconditional		Unbalanced		Balanced	
	Level	I.H.S.	Level	I.H.S.	Level	I.H.S.	Level	Log	Level	Log
Individual Treatment*During Intervention	-3.012	-0.018	-1.375	-0.063	-0.201	-0.001	2.133	-0.001	2.259	0.047
	(2.912)	(0.040)	(2.819)	(0.043)	(2.767)	(0.049)	(5.330)	(0.061)	(5.741)	(0.055)
Individual Treatment*Post Intervention	-2.150	0.040	2.174	0.025	3.725	0.136	6.929	0.079	7.120	0.111
	(3.741)	(0.052)	(4.475)	(0.078)	(4.332)	(0.109)	(7.705)	(0.090)	(8.110)	(0.088)
Group Treatment*During Intervention	3.837*	0.101**	5.213	0.117*	3.349	0.113	5.519	0.125	7.341	0.145**
	(2.268)	(0.039)	(3.178)	(0.061)	(3.651)	(0.107)	(5.173)	(0.079)	(4.710)	(0.070)
Group Treatment*Post Intervention	5.874**	0.121**	6.937*	0.184**	4.369	0.145	13.492**	0.227**	15.053**	0.260***
	(2.848)	(0.049)	(4.014)	(0.079)	(4.449)	(0.152)	(5.830)	(0.097)	(6.171)	(0.099)
Sample Size (N*T)	7299	7299	11807	11807	12537	12537	1008	1008	900	900
Number of Firms	145	145	147	147	157	157	120	120	100	100
P-value: Individual=Group During	0.058	0.033	0.087	0.016	0.423	0.315	0.622	0.199	0.440	0.259
P-value: Individual=Group Post	0.072	0.190	0.247	0.038	0.894	0.949	0.390	0.112	0.321	0.117
Control Mean in 2013	56.08	4.36	57.66	4.42	56.22	4.34	63.41	3.82	67.5	3.89
Control S.D. in 2013	51.33	0.86	45.95	0.89	51.19	0.93	57.69	0.837	59.80	0.821

Notes:

Fixed effects regressions with firm and time fixed effects. Standard errors clustered at the firm level are in parentheses.

*, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

Firm data survey is monthly employment from our firm survey; PILA data are monthly formal employment data from administrative records; EAM is annual data on total employment from the Colombian manufacturing survey.

Level denotes monthly level of employment; I.H.S. is inverse hyperbolic sine transformation.

Conditional denotes the analysis is for the group of surviving firms; unconditional codes employment as zero once a firm dies. Unbalanced is for the full set of firms matched in the EAM, balanced is for the set of 100 firms with data in all nine years from 2010 to 2018.

Table 5: Impacts on Firm Performance

	EAM Data (Annual Outcomes 2010-2018)									
	Sales		Profits		Value-Added		Production		Aggregate Index	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Winsorized Levels										
Individual Treatment*During Intervention	69	398	5	16	152	184	-208	116	0.020	0.035
	(437)	(605)	(236)	(257)	(263)	(285)	(517)	(564)	(0.115)	(0.125)
Individual Treatment*Post Intervention	697	818	42	85	92	125	380	474	0.066	0.082
	(623)	(1012)	(252)	(262)	(976)	(289)	(658)	(694)	(0.206)	(0.216)
Group Treatment*During Intervention	1088**	1354**	695***	749***	681***	727***	1210**	1295***	0.278***	0.298***
	(459)	(534)	(241)	(250)	(258)	(259)	(467)	(481)	(0.103)	(0.107)
Group Treatment*Post Intervention	1705**	1969**	647**	696**	753**	800**	1696**	1861**	0.300**	0.327**
	(676)	(766)	(309)	(331)	(326)	(349)	(684)	(737)	(0.135)	(0.145)
Balanced Panel	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size (N*T)	1008	900	1008	900	1008	900	1008	900	1008	900
Number of Firms	120	100	120	100	120	100	120	100	120	100
P-value: Individual=Group During	0.047	0.060	0.035	0.039	0.126	0.133	0.017	0.026	0.049	0.058
P-value: Individual=Group Post	0.182	0.183	0.059	0.077	0.042	0.053	0.093	0.098	0.287	0.292
Control Mean in 2017	4762	4821	1972	1998	2588	2627	4877	4946	-0.087	-0.087
Panel B: Inverse Hyperbolic Sine or Logs										
Individual Treatment*During Intervention	0.099	0.002	-0.245	0.148	-0.092	0.017	-0.141	-0.041	-0.157	0.016
	(0.094)	(0.065)	(0.218)	(0.101)	(0.126)	(0.104)	(0.097)	(0.070)	(0.255)	(0.254)
Individual Treatment*Post Intervention	0.051	0.125	-0.158	-0.042	0.117	0.207	-0.015	0.056	0.447	0.680**
	(0.107)	(0.101)	(0.182)	(0.142)	(0.146)	(0.136)	(0.111)	(0.107)	(0.325)	(0.306)
Group Treatment*During Intervention	0.250***	0.216***	0.369**	0.154	0.328***	0.337***	0.249***	0.226***	0.647*	0.824***
	(0.085)	(0.080)	(0.142)	(0.108)	(0.136)	(0.113)	(0.087)	(0.079)	(0.356)	(0.306)
Group Treatment*Post Intervention	0.264***	0.286***	0.231	0.050	0.356***	0.364***	0.248**	0.269***	1.067***	1.179***
	(0.097)	(0.101)	(0.144)	(0.139)	(0.134)	(0.138)	(0.097)	(0.101)	(0.340)	(0.363)
Balanced Panel	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Sample Size (N*T)	1008	900	962	765	1008	900	1008	900	1008	900
Number of Firms	120	100	118	85	120	100	120	100	120	100
P-value: Individual=Group During	0.009	0.027	0.268	0.970	0.036	0.056	0.005	0.010	0.084	0.056
P-value: Individual=Group Post	0.043	0.121	0.338	0.570	0.086	0.244	0.017	0.054	0.032	0.064
Control Mean in 2017	8.72	8.72	11.93	11.78	7.29	7.28	14.97	14.97	-0.847	-0.847

Notes:

Fixed effects regressions with firm and time fixed effects. Standard errors clustered at the firm level are in parentheses.

*, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

Data are from the Colombian annual manufacturing survey (EAM).

Outcomes are measured in millions of real Colombian pesos, with level outcomes winsorized at the 95th percentile.

Table 6: Potential Mechanisms for improving firm performance

	Defect rate	Inventory	Energy Costs/Sales	Labor Productivity	
	(1)	(2)	(3)	(4)	(5)
Individual Treatment*During Intervention	-0.008	-0.013	0.112	-0.091	-0.029
	(0.008)	(0.221)	(0.092)	(0.102)	(0.096)
Individual Treatment*Post Intervention	-0.008	0.208	-0.075	0.039	0.095
	(0.005)	(0.246)	(0.182)	(0.133)	(0.134)
Group Treatment*During Intervention	0.000	0.123	-0.052	0.203*	0.192
	(0.004)	(0.428)	(0.122)	(0.117)	(0.117)
Group Treatment*Post Intervention	-0.005	-0.321	0.130	0.129	0.103
	(0.005)	(0.447)	(0.092)	(0.129)	(0.137)
Balanced Panel	No	No	No	No	Yes
Sample Size (N*T)	3879	1008	1008	1008	900
Number of Firms	78	120	120	120	100
P-value: Individual=Group During	0.400	0.8102	0.3188	0.086	0.186
P-value: Individual=Group Post	0.600	0.2870	0.3159	0.497	0.953
Control Mean in 2017	0.025	13.41	3.0026	10.4	10.4

Notes:

Defect rate is the proportion of production that is faulty, and comes from firm survey collected by research team.

Remaining variables come from the EAM and are in logs or inverse-hyperbolic sine.

Fixed effects regressions with firm and time fixed effects. Standard errors clustered at the firm level are in parentheses.

*, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

ONLINE APPENDIX

Appendix 1: Examples of Products Manufactured

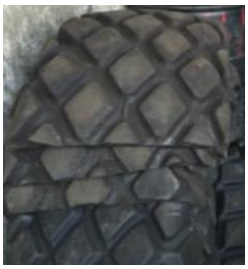
Air Filters

Glass Panels

Rubber parts

Metal parts

Plastic parts



Tires

Injection molding/cushioning

GPS tracking services

Appendix 2: Timeline

April 12, 2012: Pilot program officially launched and firms invited to apply

June 25, 2012: Deadline for firms to apply to the program

June 11, 2013: Diagnostic phase starts

October 30, 2013: Diagnostic phase ends

November 2013: Random assignment to treatment status

2013: World Management Survey administered to subsample of 72 firms with 40+ workers, as well as to random sample of 180 firms representative of Colombian manufacturing sector

March-November 2014: Individual Consulting Intervention

September 2015-April 2016: Group Consulting Intervention

November to December 2015: Round 1firm data collection (individual, group and control treatment)

January to February 2016: Round 2 of firm data collection (individual and control treatment)

March to April 2016: Round 3 of firm data collection (control treatment)

June 2016: Round 4 of firm data collection (group treatment)

November 2016: Second round of World Management Survey administered

November 2017-July 2018 : Last round of firm data collection from firms

Note: firm data collection would collect all months of data available from firm records during in-person firm visits. Timing of when this was extracted from firms varied according to CNP's contractual agreements, in which they were paid for batches of data collection at a time.

Monthly administrative data on employment are available from the PILA from January 2013 through December 2018.

Annual administrative data on sales, profits, employment, and value-added from the EAM are available from 2010 through 2018.

Appendix 3: Linking to the EAM and Selection into the Program

The *Encuesta Anual Manufacturera* (EAM) is an annual manufacturing survey conducted by Departamento Administrativo Nacional de Estadística (DANE). It is intended to cover all manufacturing firms with 10 or more workers. We supplied the list of our firms and their tax identification numbers (NIT) to the DANE team, who matched our experimental firms to their dataset.

The DANE team were able to match 120 of the 159 experimental firms (43 control, 37 individual, 40 group) to the EAM, including 114 firms in 2012 and 2013. They noted that the non-matched firms included companies that trade in auto-parts or that are multi-activity, and so did not meet the eligibility criteria for the EAM, as well as some firms that should have been eligible but were not able to be located. Table A3.1 compares our baseline characteristics for the matched and unmatched firms, and finds that the unmatched firms were smaller, young, and less likely to be exporting than matched firms.

Table A3.1: Comparison of Matched and Unmatched firms in EAM

	Mean for Matched Firms	Mean for Unmatched Firms	P-value
Number of Employees	69.5	28.8	0.000
Small	0.5	0.8	0.000
Medium	0.5	0.2	0.000
Cundinamarca	0.5	0.4	0.441
Valle	0.2	0.2	0.791
Labor Productivity	31.8	27.8	0.238
Finance	51.8	47.4	0.083
Human Resources	43.9	38.5	0.015
Logistics	47.4	42.0	0.023
Marketing	46.7	42.0	0.086
Production	48.3	41.0	0.002
Level 2 Supplier	0.9	1.0	0.257
Metal Products	0.6	0.6	0.739
Plastic Products	0.2	0.2	0.649
Firm Age	27.1	15.4	0.000
At least one overseas customer	0.5	0.2	0.003
Sample Size	120	39	

The EAM can then be used to examine selection into the management improvement program. The matched firms in 2012 are found in 35 different 4-digit ISIC industry codes. There are 3,406 firms in the EAM in these 35 industries, so our sample constitutes only 3.3 percent of all firms. However, these industries include many sectors that supply inputs to many types of products (e.g. “manufacture of basic forms of plastic”, “manufacture of iron and steel”, “manufacture from basic rubber”, “manufacture from glass”), with the majority of firms outside of the experiment in these sectors likely to not be making autoparts. The most common 4-digit ISIC code in our sample is 2930 “manufacture of parts, pieces, and accessories for automobiles”. Our experimental sample constitutes 22 of the 109 firms in the EAM with this code (20% of all firms), 22 out of 92 of these firms with 10 to 250 workers (24%), and 13 out of 38 firms with 50 to 250 workers (34%), and 12 out of 38 firms with 25 to 100 workers (32%).

Table A3.2 compares baseline means for those in the management improvement program to the other firms in the EAM for the full sample of 35 industries, and for the sub-sample of firms in the ISIC code 2930. We see that the average number of employees are not statistically different between those in the program and those not in the program, but the program firms are more likely to be medium sized (50 or more workers) and less likely to be small firms. We cannot reject that sales per worker and production per worker are the same on average for firms in the program and those not in the program.

Figure A3.1 provides a histogram of baseline employment for firms in the program in this sector to those not in the program. We see that the smallest and largest firms in this industry did not take part in the program.

Figure A3.1: Histogram of Number of Employees in 2012 of Autoparts Firms (ISIC code 2930) in Management Improvement Program compared to other firms in EAM in this sector

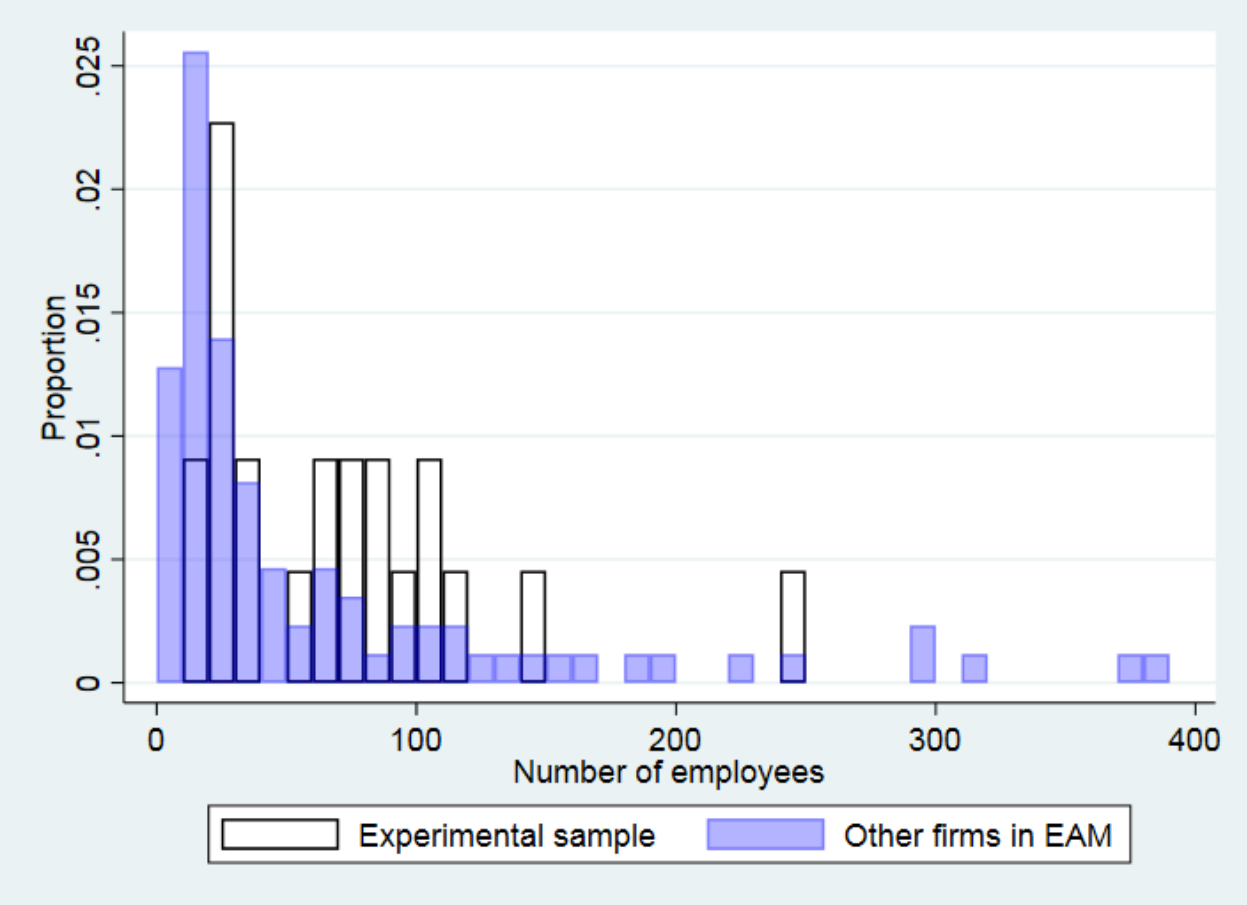


Table A3.2: Selection into the Management Extension Program

	All industries that at least one program firm is in:			ISIC Code 2930: Autoparts		
	Experimental Firms in EAM	Other Firms in EAM	p-value	Experimental Firms in EAM	Other Firms in EAM	p-value
Number of Employees	67	60	0.58	69	83	0.72
Small	0.5	0.72	0.00	0.41	0.6	0.46
Medium	0.5	0.28	0.00	0.59	0.4	0.46
Cundinamarca	0.49	0.5	0.79	0.91	0.7	0.02
Valle	0.19	0.12	0.05	0	0.05	0.31
Sales per worker	83615	149165	0.38	65270	108861	0.21
Production per worker	83422	160393	0.48	63334	109466	0.19
Sample size	114	3292		22	87	

Appendix 4: Data Appendix

A4.A. Management practices indicators:

The 141 management practices defined by CNP can be divided into five main areas: Finance, Production, Logistics, HR, Marketing. Each of these areas can be itself divided into five to eight sub-areas. The score of the five main areas is the average of the score of their sub-areas. Below we discuss each of these sub-areas and explain which practices were considered to calculate their score. At the most basic level, each single practice is graded on the following scale: 1 = "Not existing", 2 = "In construction", 3 = "Formalized", 4 = "Implemented", 5 = "Operating under control". For some indicators, the 1 to 5 scale does not exactly refer to the implementation stage of a practice, instead it indicates how developed or optimized a specific aspect is – for instance whether strategical goals and individual responsibilities are clear to each worker. Such information was collected in three stages: during the diagnostic phase, during the intervention, and once a year after the intervention.

Human Resources

i. Strategic objectives leverage on people's talent

The first aspect of Human Resources relates to the alignment of employees' objectives with corporate strategy, and to the clarity of such objectives for each employee. Here we consider four components. The first one evaluates how strategic objectives leverage on people's and teams' talent. The second component assesses whether there are human talent development plans, and whether these leverage on corporate strategy. The third component assesses whether a strategic plan is defined, that includes clear objectives and goals concerning human talent. The last component assesses whether the skill development plans are defined also for the operational level.

ii. Competency-based management model for human talent development

The focus of this measure is on whether the company manages employee competences – based on the business strategy – in order to develop human talent. It is comprised of two measures. The first one assesses whether human resources are monitored based on their impact on the strategic objectives of the organization. The second component addresses the development of work profiles, which must be defined and aligned with business competencies.

iii. Organizational structure prepared to contribute to the achievement of strategic goals

The third sub-area evaluates whether the formal and informal structure of the organization allows the realization of corporate strategy. Is there a formally defined structure? Are all roles well defined at every level of the organization? Three measures are taken into consideration. The first one evaluates if the management's focus is on processes which are aligned with the strategy of the firm. The second one assesses whether a communication system between the different processes of the organization has been developed. The last measure assesses whether a communication system between the different levels of the organization has been developed.

iv. Program of human talent development (according business competences)

This measure evaluates how the organization works on building and retaining human talent to achieve a competitive advantage over the competition. Two components are considered: Management of

development plans (career plans) for employees at managerial level, and the level of application of the sector's technical norms for the development of technical operational competences.

v. Organizational climate

The focus of this sub-area is the management of a work climate. Work climate must be appropriate for the development of Human Capital and directed towards the achievement of corporate strategy. We consider three components. Is there a culture of monitoring work climate, as strategic lever? Are there programs to improve work climate? At which level are risks for health and safety controlled?

vi. Social responsibility within the enterprise

Here we evaluate how the company manages its internal social responsibilities. This measure is comprised of three components. The first one assesses whether there are programs of improvement of the family environment of employees, in order to incentivize their productivity. The second one verifies whether a formal contracting system is in place, which generates wellbeing and productivity in workers. The last one evaluates the implementation of a system of recognition and retribution of new ideas and improvement suggestions at the operational level.

vii. Promotion of an open-communication/high-performance organizational culture, and of a culture of high personal involvement

Three measures are considered for this indicator. Did the company develop a culture of control and periodic monitoring of result achievement? How developed is the performance-based reward system for the management? How developed is the performance-based reward system for employees at the operational level?

Production

i. Alignment of functions at the operational, managerial and directive level

The first sub-area of Production focus on whether all people working in the plant know the corporate strategy and work to realize it. To achieve this, it is necessary that all workers and processes have improvement goals aligned with corporate strategy. This measure is comprised of five components. The first two evaluates the implementation and monthly monitoring of strategic goals between the Plant Manager and his/her supervisor. The third and fourth components assess whether strategical goals and individual responsibilities are clear to each worker, and whether each worker has improvement goals. The last component assesses whether the performance of teams at the operational level is evaluated based on the strategic goals.

ii. Definitions and management of the most important operational processes

Here we evaluate how operational processes are defined and managed, from the order to the delivery of the final product. Do they allow to accomplish the strategy (Standards, Policies, Roles, 5s, Layout, Established Processes)? This sub-area includes six components. The first one evaluates whether processes are well identified and have a proper description (VSN, SIPOC). The second one assesses whether the plant layout allows optimal material flow. The third one concerns the implementation of a 5S program in the plant. The fourth one evaluates how bottle necks are identified and managed. The last two components evaluate standards, specifications and work instructions used by workers, and how these are verified by supervisors.

iii. Formal method to measure and manage the plant's efficiency (Waste, Hours paid/Service capacity, machinery's efficiency)

The third sub-area evaluates how the company measures and manages the main KPIs of the plant, such as team efficiency, efficiency in the use of material, response time, etc. The first of components of this sub-area concerns the monthly measure of the plant's KPIs (OEE, Waste, Defects, Lead time, Others). The second indicator concerns weekly or bi-weekly management of KPIs' goals (OEE, Waste, Defects, Lead Time, Others). The third one assesses whether improvement programs for KPIs (times and quality) are developed applying instruments of plant management. The last one assesses whether a culture of daily recollection of facts and data is in place, in order to demonstrate improvement in processes.

iv. Recollection of information regarding results, continual improvement, and performance of processes

Here we assess how the company is managing data and information regarding processes, results and continuous improvement. The four components of this sub-area are the following: Is there a culture of visual management with daily-updated graphs of machinery performance? Are duration and quality of each process recorded daily by the responsible worker? Does the Administrative Management make sure that monitoring instruments are in good condition and precise? Is there a monitoring and sampling plan to capture the information necessary to the improvement of processes?

v. Process to detect and solve anomalies in the execution of tasks

The focus of this sub-area is to evaluate how anomalies in processes are managed within the plant. It is comprised of five components. The first one assesses whether there is a mechanism so that workers report anomalies of time and quality to their supervisors. The second one assesses whether criteria are defined to realize analysis of anomalies. The third one concerns the daily analysis of time and quality anomalies by supervisors and workers. The fourth one assesses whether supervisors and workers manage improvement plans to eliminate time and quality anomalies. The last component concerns job descriptions, and whether they include responsibilities of anomalies solving.

vi. Technical planning of production based on the analysis of demand

The focus of sixth sub-area is the planning of production. Is such planning based on a statistical analysis of clients' orders? Does such planning guarantee the flexibility necessary to achieve a high level of service? Four components constitute this sub-area. The first one assesses whether meetings to revise programming take place between production and sales areas. The second component evaluates the use of statistic methods to collect information and analyze production programming, according to demand variation. The third one evaluates production planning to ensure the availability of material for the monthly, weekly and daily program. The last component evaluates monitoring and management of service to clients (deliveries in quality, time and quantity).

vii. Management of safety during the process, contingencies, emergencies / impact on the environment

Here we assess how the company monitors its impact on people and environment, which actions are undertaken to mitigate any negative impact, and how it complies with safety and environmental norms and regulations. This sub-area is comprised by five measures. The first one concerns the compliance with safety requirements, laws and norms. The second measure assesses whether the necessary norms and standards of safety within the plant are well defined. The third one evaluates the management of the indicators of industrial safety within the plant (number of accidents, level of noise, temperature). The fourth one concerns

monitoring and management of the plant's environmental impact. The last measure assesses compliance with the norms regarding evacuation routes and cleared zones for fire-fighting equipment.

viii. Maintenance guarantees the optimal condition of infrastructure

The last sub-area of Production evaluates the maintenance plan, how maintenance is monitored and managed and how maintenance is related to the creation of value by the enterprise. All this is paramount to guarantee optimal condition of machinery, furniture, equipment and tools. This measure reflects the following four points. Is there a preventive maintenance plan for the equipment? Are technicians able to rapidly repair damage to the machines? Are replacements available, so to allow to rapidly repair damage to the machines? Does Maintenance Management work with indicators such as MTTR, MTBF, Availability?

Logistics

i. Process of alignment of functions at the operational, managerial and directive level

The first sub-area of Logistics looks at the alignment of functions, and at the deployment of the organizational strategy. It is comprised of three components. The first one concerns the implementation of strategic goals between the Logistics Head and his/her supervisor, and whether there are specific projects to achieve such goals. The second component assesses whether there is a monthly control of strategic goals by the Plant Manager and the supervisor. The last component concerns the alignment of employees' objectives in the logistics area with the firm's strategic goals.

ii. Structure and management of the supply chain (planning, purchases and provisions, storage raw material, plant supply, storage finished product, distribution, client service)

Here we evaluate if employees in the logistics area understand their roles and activities. In this sub-area there are four measures. The first one evaluates procedures and work instructions for logistics processes. The second measure is concerned with the layout of the areas of logistic operations in the supply chain. The third component assesses if a 5S plan for the supply chain is in place. The last component evaluates monitoring and management of KPIs in the logistic process (inventory, lead time, service level).

iii. Planning and management of demand / alignment of productive and logistic processes

This sub-area evaluates the procedure through which demand is planned and the reaction to changes in the established plan. Here we have four distinct components. The first one assesses whether a statistical system is in place, in order to study and analyze demand. The second component concerns the definition of the demand's planning, and whether such definition is updated with annual, trimestral and monthly frequency. The third component evaluates whether communication between logistics and the areas of marketing and sales goes through a system that includes rules to change the production plan. The last component evaluates the way a firm monitors and manages the compliance with the budgets of production planning.

iv. Planning, management and control of inventories of raw material, supplies, product on process and finished product (Inventory Policies)

This sub-area evaluates the design of the inventory system, and the maintenance of inventory levels. The five components upon which this measure is based are the following. The first one assesses whether the levels of inventory (raw material, semi-finalized product WIP, finished product) are kept at an optimal level

related to the variation in demand. The second component assesses whether the inventory movement it is recorded daily and controlled weekly. The third component states whether a methodology of classification of inventory ABC is in place, in order to establish policies of inventory, supply, storage and control accordingly. The fourth component verifies the use of MRP systems, where product structures are defined, in ways that allow to plan the material needed to comply with production orders. The last component evaluates whether processes are in place, so to guarantee the rotation of inventory according to "First in, first out" schemes.

v. Supply system

This sub-area concerns the relation with suppliers, the way in which suppliers are evaluated, and the control the firm has over realized purchases. It is comprised of five measures. The first one concerns the management of policies and processes for the selection and evaluation of suppliers. The second measure concerns the management of suppliers' development. The third measure focusses on the management of raw material prices and supplies. The fourth measure assesses whether Lead Time of suppliers is managed and taken into account in the planning of material supply. The last measure assesses whether purchased items are verified in terms of quantity, quality and opportunity of delivery.

vi. Storage system

Five components are taken into account while evaluating the storage system. The first one is the management of the inventory of obsolete and non-compliant products. The second one is the implementation of a system to administrate storage locations (layout and 5S). The third one evaluates the implementation of industrial security norms in the warehouse's operations. The fourth one concerns the use of standards and procedures in the storage operations (picking and packing). The last component evaluates the monitoring and improvement of the storage operation time (picking and packing).

vii. Distribution system

This last sub-area of Logistics concerns the delivery of the created value to the client. It is comprised of four components. The first one evaluates efficiency in the processes of loading and unloading. The second one evaluates monitoring and management of the efficiency in the delivery process (perfect deliveries). The third component concerns the management of transport routes to reduce costs. The fourth component evaluates the management of reverse logistics for those products, materials or supplies that have to return to the company's premises. The last component evaluates whether the management of distribution takes into account the current legislation regarding freight transit.

Marketing

i. Elaboration, management and control of the marketing plan

This measure evaluates the design of the guiding document of commercial activities and its alignment to the organization's strategy. Such indicator is comprised of seven components. The first two assess the implementation of an analysis of trends (economic, commercial, technological, political and social) and of risks (e.g. free commerce, supply, variations in exchange rate, infrastructure, etc.). The third indicator evaluates the segmentation of products, technology, clients, consumers, etc. The fourth component assesses whether commercial strategies are based on contribution margins. The fifth component evaluates the alignment of the marketing and sales plan with Business Strategy. The sixth indicator assesses whether price, promotion and growth policies are defined using the contribution margins. The last indicator addresses monitoring of sale behavior and trends, and of changes in the marketing plan.

ii. Processes of market research

This measure indicates how the company conducts market research, and is composed by three components. The first one addresses if and how the company conducts inquiries with clients and potential clients. The second one assesses whether the company conducts periodic monitoring of competitors' offers. The last component evaluates if and how the company conducts research of marketers and/or distributors.

iii. Client and after sales service

This measure evaluates the company's approach to client satisfaction and is comprised of four measures. The first one evaluates the management of clients' complaints and requests. The second measure concerns the analysis of products' performance in the market. The third measure assesses whether in the company there is a culture of continuous improvement of products and services. The last component verifies if the company holds periodic meetings to discuss clients' feedback.

iv. Sales management

This sub-area focusses on the elaboration, management and control of the sales plan. We consider five indicators. The first three assesses whether the company is holding three different type of meetings: with the distribution channels (to capitalize opportunities in the market), planning meetings between sales and production, and meetings of the sales group to analyze sales behavior and trends. The fourth component assesses whether periodic training of the sales team takes place. The last indicator states whether sales agents are evaluated based on performance.

v. Relationship management

This measure is built on three components evaluating whether the company conducts three types of evaluation studies: of its cooperation with suppliers, of its cooperation with clients, and of its cooperation with competitors.

Finance

i. Alignment of the financial process with corporate strategy

Four components indicate whether strategic objectives and goals are clear at all levels of the financial process, and whether everyone is committed to such goals. The first component refers to the alignment of the Financial Head and Deputy Head with corporate strategic goals. The second component indicates whether a system of monitoring and control of financial goals and objectives is in place. The third indicator refers to the frequency in which financial objectives and goals are achieved. The last component evaluates the financial support to the management processes of the organization.

ii. Structure of the administrative and operational information system

The administrative information system is evaluated based on monitoring and controlling of processes, in its effectiveness of analysis and decision making. This is reflected in five measures. The first measure evaluates the structure of the corporative information system. The second one assesses whether the setup of administrative and operational business' information is appropriate. The third one states if Product Structures are associated with cost and profitability margins (standard, estimated, reals). A fourth indicator refers to the protection of the corporative information system, whereas the last one evaluates the organization of the corporative information system.

iii. Formulation and management of budgets

This sub-area evaluates how the firm formulates and manages budgets. The measure is comprised of four components. The first two focus on the existence of a Master Budget (operational, financial and of investment) and on its control and monitoring (agendas, finances, investment). The third component assesses Tax Planning, and the last one evaluates how deviations from Master Budget are analyzed (regarding costs, expenses, sales, working capital, investment).

iv. Financial management of results

The fourth component of Finance reflects how well the company monitors and manages indicators of financial management, and how it analyses them to undertake corrective action. Three components build this measure: the first evaluates the structure of control and monitoring indicators (KPIs), the second one the agenda of financial management meetings, and the third one how working capital is managed.

v. Programs of financial improvement (costs and expenses, working capital, investment)

This sub-area evaluates how projections and saving goals are realized. It is comprised of three components answering the following three questions: is there a program of efficient administration of costs and expenses? Is there an action plan for the compliance with financial improvement programs? Is the available financial information appropriate?

vi. Analysis and management of investment projects

This sub-area evaluates the process which the firm uses to plan, realize and follow up the purchase of fixed assets. This measure is made of three components. The first component assesses if a program of calculation of investment projects exists and if it is aligned with strategy. The second one verifies whether there is a policy regarding capital investment (CAPEX) and other smaller investments. The last one concerns the implementation of cost-benefit analysis for the different projects and firm's investments.

vii. Information systems

The second-last sub-area of finance evaluates if the information systems are interrelated and if strategies are in place to safely conserve information. Three aspects are considered here: the recollection and storage structure of the administrative information system, recollection and storage structure of the operational information system, and validation of information.

viii. Structure of the costing system

The last sub-area of finance evaluates whether the costing system supplies real and updated information, so to identify cost anomalies in any process. The first of four components reflects the implementation of a costing systems. The second component assesses if results (value estimates and real) are being validated. The last two components evaluate absorption capacity of installed structure and workforce efficiency.

A4.B Key Performance indicators collected directly from firms

Collecting key performance indicators directly from firms was complicated due to several factors. First, a consequence of poor management is that firms did not routinely and consistently keep records of some KPIs. Firms would sometimes change the units of measurement at times from pesos to physical units, and the type of physical unit they used (e.g. from number of items to kilograms).[19] Second, data collection in the firms was conducted during on-site visits by CNP. We hired Innovations for Poverty Action to provide an independent check on this data, and to help in extracting data from the firms – this included oversight of both the management practice data and the KPI data. But CNP had breaks in its contracts, which meant data collection halted for months at a time, and they had a long list of KPIs they wanted from firms, which increased the burden on firms of reporting. The long length of our project and follow-up period also meant that some firms who initially cooperated refused to provide data after several years. The result was that some firms dropped out of providing follow-up information, even after repeated follow-up visits seeking just a few key variables. Third, ten of the firms closed during the course of the study (4 control, 3 individual treatment, and 3 group treatment, p-value of equality of death rates 0.911).

These three factors mean that from our surveys we only have both employment and sales data through to December 2017 for 105 firms (69% of the sample), comprising 33 control firms, 37 individual treatment firms, and 35 group treatment firms (p-value of equality of attrition rates is 0.744). Table A5.2 compares the baseline characteristics of these firms to those that attrit, and shows that we cannot reject equality of means. Moreover, balance on baseline observables for those firms which do report is similar to our balance on the overall sample. Nevertheless, we use firm fixed effects in our estimation of impacts on firm outcomes to control further for any time-invariant differences among firms.

We use the following three variables, each recorded monthly.

Defect rate: this is defined as the ratio of faulty production to total production. Faulty production is defined as not in condition to be sold, and is determined by the firm. There are several key measurement issues with this measure. First, firms vary in whether they record production in physical units (e.g. number of items, kilograms) or in pesos. Secondly, some firms would calculate this product only for a specific production line or product, and not for the whole plant. Thirdly, in a few cases, firms changed the way they measured these units over time. IPA and CNP worked together to identify these cases, and the series we use is for the set of firms with a consistent measure.

Net sales: Total sales (gross sales) minus devolutions (discounts, etc.). This is taken directly from the Profit & Loss Statement (P&L) or records of the firms. Given the gaps in coverage, we use this only in Appendix 10. We deflate by the Colombian PPI to express this in millions of 2017 real pesos.

Total employees: All employees of the firm which are considered "stable or long term", independently of the contract type. There are no standard criteria to define what a "long term"

[19] These changes in units also occurred because firms would produce different products at different times, depending on what orders they received.

employee is. This is defined by each firm. They calculate it considering the totality of the firm. Our administrative measures include all workers for whom the firm enters into the PILA.

A4.C PILA and EAM Variables

We use the following outcome variables from the PILA (Unified Register of Contributions) and EAM (Annual Manufacturing Survey):

PILA:

PILA Employment: the total number of workers employed in a firm in a given month who are registered for social security.

EAM:

Total Employment: Total employment in the firm, measured as the sum of permanent employees, owners, directly employed temporary employees, and temporary employees indirectly hired through third-party labor contractors. Firms are asked to report the average employment used in the reference year.

We convert the following variables into 2017 real pesos using the Colombian Manufacuring Producers Price Index from DANE:

Total sales: Annual sales, measured in millions of pesos.

Annual profits: Value-added less wage costs, measured in millions of pesos.

Value-added: Total value-added, measured in millions of pesos. DANE calculates this as the difference between the value of gross production and the amount spent on consumption of intermediate inputs.

Production: The value of annual production, measured in millions of pesos.

Aggregate Performance Index: The average of standardized z-scores of total sales, annual profits, value-added and production. For each variable, a z-score is calculated by subtracting the control mean from 2012 and dividing by the control standard deviation from 2012.

Labor productivity: Total value-added per worker, measured in thousands of pesos.

Inventories: Total value of inventories, measured in thousands of pesos, as of December 31 of the reference year.

Energy costs/Sales: the ratio of annual energy expenditure to annual sales

Appendix 5: Drop-Out and Attrition

Table A5.1 shows that the firms that completed the interventions are similar on baseline characteristics to those which dropped out.

Table A5.1: Comparison of Baseline Characteristics of Firms that Completed Interventions to Drop-Outs

	Individual Treatment	Group Treatment

	Completed	Dropped Out	p-value	Completed	Dropped Out	p-value
Number of Employees	62.2	54.4	0.746	52.9	53.1	0.981
Small Firm (<=50 employees)	0.59	0.57	0.940	0.58	0.59	0.974
Medium Firm (>50 employees)	0.41	0.43	0.940	0.42	0.41	0.974
Cundinamarca	0.54	0.14	0.049	0.42	0.35	0.665
Valle	0.09	0.14	0.645	0.25	0.18	0.559
Labor Productivity	32	30	0.780	32	39	0.278
Financing Practices	48	47	0.820	53	52	0.855
Human Resources Practices	42	40	0.625	44	43	0.784
Logistics Practices	43	43	0.911	49	43	0.175
Marketing Practices	43	43	0.920	46	46	0.948
Production Practices	46	52	0.296	47	44	0.371
Level 2 Supplier	0.93	1.00	0.496	0.92	0.94	0.758
Metal Products	0.50	0.57	0.731	0.47	0.65	0.242
Plastic Products	0.15	0.29	0.390	0.19	0.24	0.738
Firm Age (Years)	23.3	21.8	0.829	20.9	24.6	0.375
Anexo K score	44	45	0.905	48	46	0.487
USD Sales in 2013	2688709	6424375	0.189	2355771	2101746	0.799
Export at all in 2013	0.43	0.29	0.465	0.47	0.41	0.687
Sample Size	46	7		36	17	

Table A5.2 compares the characteristics of those firms for which we have December 2017 sales and employment data from our surveys to the attritors, and then shows the sample of non-attritors is reasonably well balanced on baseline characteristics.

Table A5.2: Comparison of Baseline Characteristics of Non-Attritors to Attritors, and Balance on Non-Attiting Sample

	Full Sample			Sample of Non-Attritors			
	Non-Attritors	Attritors	p-value	Control	Individual	Group	p-value
Number of Employees	58.9	59.8	0.921	54.9	68.2	52.9	0.441
Small Firm (<=50 employees)	0.58	0.61	0.716	0.67	0.51	0.57	0.426
Medium Firm (>50 employees)	0.42	0.39	0.716	0.33	0.49	0.43	0.426
Cundinamarca	0.50	0.43	0.349	0.58	0.51	0.43	0.480
Valle	0.16	0.17	0.939	0.18	0.08	0.23	0.174
Labor Productivity	30	32	0.460	26	32	32	0.054
Financing Practices	51	50	0.810	51	48	53	0.154
Human Resources Practices	44	40	0.059	45	43	44	0.906
Logistics Practices	47	44	0.145	50	44	48	0.106
Marketing Practices	46	44	0.261	47	45	47	0.841
Production Practices	47	46	0.478	47	48	46	0.867
Level 2 Supplier	0.94	0.93	0.679	0.94	0.95	0.94	0.993
Metal Products	0.57	0.65	0.353	0.79	0.46	0.49	0.004
Plastic Products	0.15	0.22	0.276	0.09	0.16	0.20	0.404
Firm Age (Years)	24.1	24.1	0.997	27.6	24.6	20.2	0.085
Anexo K score	47	45	0.173	48	46	48	0.538
USD Sales in 2013	2449562	1917141	0.342	1739554	2991799	2564553	0.133
Export at all in 2013	0.47	0.41	0.480	0.48	0.46	0.46	0.969
Sample Size	105	54		33	37	35	

Notes: Attrition defined as not having firm sales and employment data reported from firm records in December 2017. This can arise from firms refusing to provide this information, as well as from firm death. P-value in column 3 is for a t-test of equality of means by attrition status.

Columns 4 through 6 provide baseline means by treatment status for the sample of non-attritors. P-value in column 7 is for F-test of equality of means.

Table A5.3 shows that firm survival rates are high and similar across treatment groups:

Table A5.3: Impact on Firm Death

	Survives in full sample	Survives in PILA data	Survives in EAM data
Individual Treatment	0.019	0.020	-0.0246
	(0.049)	(0.051)	(0.0683)
Group Treatment	0.019	0.020	0.0141
	(0.049)	(0.051)	(0.0663)
Survival measured at:	mid-2017	Dec 2018	Year 2018
Sample Size	159	157	115
Control Mean	0.937	0.936	0.907

Note: robust standard errors in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

Appendix 6: Robustness of Management Improvements to Aggregation Weights and Sample Attrition

Robustness of Management Impacts to Choice of Aggregation Weights

Our measures of management practices are averages of the different practices. The Anexo K overall index is an average of the 35 sub-indices, and ranges from 20 (indicating scores of 1 for every individual practice) to 100 (indicating scores of 5 for every individual practice). With any aggregate index, there is always a question as to the appropriate choice of weights, and of how sensitive the results are to alternative weighting schemes.

Table A6.1 examines robustness to different choices of how to aggregate the 141 practices. Column 1 shows our aggregate index from Table 2. Columns 2 through 5 then consider four alternative weighting schemes. Column 2 uses the first principal component of the 141 practices; Columns 3 and 4 use lasso regression to identify the sub-set of practices which best predicts baseline log employment and labor productivity respectively, and then post-lasso regression to form the weights. This chooses 19 practices to weight according to their predictive power for employment, and 14 to weight for their predictive power for labor productivity. Finally, column 5 uses the subset of firms for which we also have baseline data from the World Management Survey, and uses lasso to choose weights that best predict the baseline WMS score, which selects only 6

practices.[20] The coefficients cannot be directly compared across columns in terms of magnitudes, but can be considered relative to the control group standard deviation. The estimated treatment effects are 0.8 to 0.9 standard deviations (s.d.) when using our aggregate index, 0.9 to 1.0 s.d. when using principal components, 0.6 s.d. when weighting to predict employment, 0.8 s.d. when weighting to predict labor productivity, and 0.7 to 1.1 s.d. when weighting to predict the WMS score. Thus, regardless of the choice of weights, we find the treatment impacts are positive, similar in magnitude, and statistically significant.

[20] The smaller number of practices chosen is likely because of the much smaller sample for which the WMS is available.

Table A6.1: Robustness of Impact on Management Practices to different weighting schemes

	Overall Anexo K	Principal component	Lasso Log Employ.	Lasso Productivity	Lasso WMS
Panel A: Unbalanced Panel					
Individual Treatment*During Intervention	9.703***	6.014***	0.227***	7.065***	0.079**
	(1.370)	(0.946)	(0.085)	(1.238)	(0.036)
Individual Treatment*Post Intervention	9.620***	6.012***	0.286**	8.297***	0.140***
	(1.830)	(1.217)	(0.115)	(1.811)	(0.041)
Group Treatment*During Intervention	11.971***	7.266***	0.403***	9.269***	0.240***
	(1.660)	(1.177)	(0.090)	(1.463)	(0.040)
Group Treatment*Post Intervention	8.544***	5.512***	0.301***	7.596***	0.225***
	(1.894)	(1.220)	(0.106)	(1.706)	(0.040)
Sample Size	225	200	213	217	221
P-value: Individual=Group During	0.145	0.208	0.020	0.111	0.000
P-value: Individual=Group Post	0.533	0.658	0.862	0.670	0.043
Control Mean	55.98	5.59	2.46	43.01	0.93
Control SD	10.79	6.03	0.47	9.66	0.20
Panel B: Balanced Panel					
Individual Treatment*During Intervention	9.861***	6.048***	0.273**	7.302***	0.100**
	(1.756)	(1.327)	(0.119)	(1.602)	(0.049)
Individual Treatment*Post Intervention	9.757***	5.972***	0.309**	8.451***	0.148***
	(2.014)	(1.402)	(0.122)	(2.003)	(0.044)
Group Treatment*During Intervention	12.118***	7.494***	0.445***	9.624***	0.263***
	(2.029)	(1.525)	(0.118)	(1.781)	(0.051)
Group Treatment*Post Intervention	8.889***	5.736***	0.361***	8.009***	0.242***
	(2.067)	(1.416)	(0.111)	(1.914)	(0.043)
Sample Size	202	178	190	194	198
P-value: Individual=Group During	0.152	0.174	0.032	0.114	0.000
P-value: Individual=Group Post	0.627	0.844	0.539	0.797	0.032
Control Mean	55.98	5.59	2.46	43.01	0.93
Control SD	10.79	6.03	0.47	9.66	0.20

Notes:
Panel A is for the 124 firms for which Anexo K management practices are measured post-baseline, panel B for the 101 firms for which practices are measured both during and after intervention.
Robust standard errors in parentheses, clustered at the firm level. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively.
Anexo K management practices are 141 management practices divided into five sub-areas.
Ancova estimation controls for baseline (December 2013) mean, time and triplet fixed effects.
Principal Component takes the first principal component of the 141 practices.
Remaining columns using Lasso to choose the subset of practices that best predict log baseline employment, log labor productivity, and the WMS baseline management score respectively, with post-Lasso coefficients then providing the weightings on the different practices used.

Robustness of Management Impacts to Sample Attrition

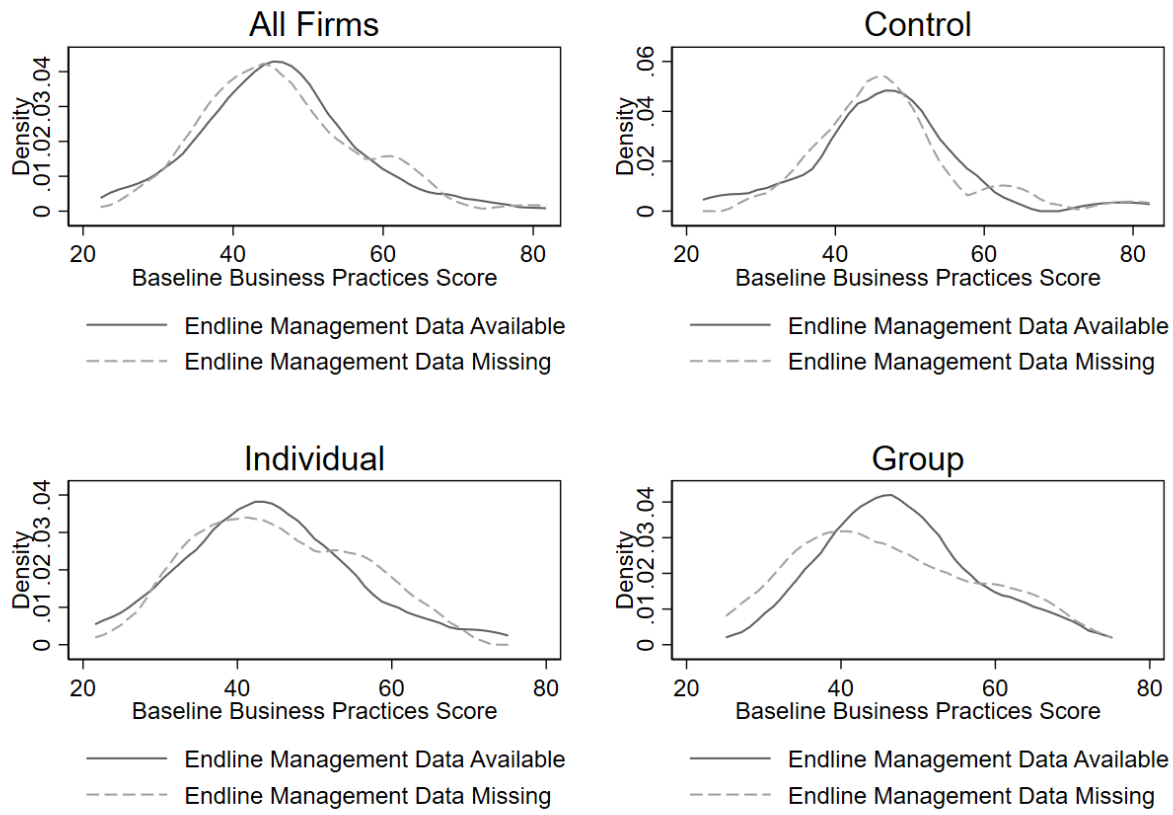
Table A6.2 shows the availability of our management score data by time period and measure. The greatest data availability is for the Anexo K measure, but this still suffers from attrition, while the WMS and MOPS data are available for subsets of the same only.

Table A6.2: Management Data availability by measure and time period

		# Firms with Data by Treatment			
Measure	**Period**	**Control**	**Individual**	**Group**	**Data source**
Anexo K management score	2013	53	53	53	Anexo K collected by CNP
	2014	42	46	0	Anexo K collected by CNP
	2015	26	40	37	Anexo K collected by CNP
	2016	0	0	36	Anexo K collected by CNP
WMS management score	2013	26	24	27	WMS collected by LSE
	2016	20	19	31	WMS collected by IPA
MOPS management data	2012	28	34	33	Collected retrospectively by IPA
	2017	28	34	33	Collected by IPA

Figure A6.1 compares the distribution of baseline management practice data for firms which attrit and do not have endline (2015 for the control and individual treatment, 2016 for the group treatment) Anexo K data. We see that the distribution of those with and without follow-up management data is similar, both for the full sample, and when we split by treatment status. We cannot reject equality of distributions between attritors and non-attritors using a Kolmogorov-Smirnov test of equality of distributions. This shows that attrition is not selective on initial management practices.

Figure A6.1: Distribution of Baseline Anexo K Management Practices by Whether or Not Endline Management Data are Missing



Notes: Kolmogorov-Smirnov tests of equality of distributions of baseline management practices between firms with missing endline management data and firms with endline management data have p-values 0.975 (all firms), 0.990 (control firms), 0.964 (individual treatment), and 0.425 (group treatment).

Note that our main estimates of the treatment effect are for a balanced panel, and include randomization triplet fixed effects. Coupled with the above analysis which shows no selection on baseline management practices into having follow-up data, and Figure 2 which shows clearly the change in distribution of practices for this balanced panel, this suggests our main results are not being driven by selective attrition. Nevertheless, as a further sensitivity check, Table A6.3 provides Lee bounds for the treatment impacts. Table A6.2 shows we have substantially more control firms reporting management practices in 2014 than 2015, so less trimming is required when estimating the impact during the year of intervention than for the post-intervention impact. We see that both the treatments have significant impacts even at the lower bound for the during intervention period. In contrast, the bounds become wider for the post-intervention period. If all the additional firms that attrited from the control group were the best managed firms, then we could not conclude the intervention had had a positive effect. We can examine this assumption using the control firms that attrited between 2014 and 2015. The 16 control firms that attrited had first follow-up (2014) Anexo K scores with a mean of 51.4, while the 26 control firms that did not attrit had 2014 mean Anexo

K scores with a mean of 52.8 (p-value 0.72). Thus, not only is there no evidence of selective attrition on baseline management practices, neither is there evidence of endline selective attrition based on first follow-up management practices. This strongly suggests that the assumption that it was all the best-managed firms in the control group that differentially attrited is very unlikely to hold, so that the Lee lower bound is unlikely to be applicable.

Table A6.3: Lee Bounds of Impact on Anexo K Score

	Individual Treatment Effect	Group Treatment Effect
Impact during intervention		
Lee lower bound	6.303**	9.368***
	(2.723)	(3.290)
Lee Upper bound	9.746***	16.610***
	(3.065)	(2.851)
Impact post-intervention		
Lee lower bound	1.076	4.784
	(3.628)	(3.218)
Lee Upper bound	13.993***	13.913***
	(3.011)	(3.158)
Sample Size	106	106
Proportion trimmed		
for during intervention	8.7%	16.7%
for post-intervention	35.0%	27.8%

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

Appendix 7: Impacts on Individual Management Practices

Table A7.1 shows the breakdown of significant improvements in management practices within the Anexo K index:

Table A7.1: Summary of Impacts at the Sub-Index and Individual Practice Level

	Sub-Indices			**Individual Practices**		
	#	# sig. Ind.	# sig. Group	#	# sig. Ind.	# sig. Group
Finances	8	6	5	29	16	15
HR	7	3	2	20	11	6
Logistics	7	5	2	31	7	9
Marketing	5	3	3	22	9	13
Production	8	6	8	39	22	30
TOTAL	35	23	20	141	65	73

Note: lists number of practices that are statistically significant at the 5% level post-intervention.

Table A7.2 details the individual management practices that have treatment effects of 0.8 or more (on a 5-point scale).

Table A7.2: Practices that increase by 0.8 or more from at least on treatment

		Treatment Effect		P-value
Practice Area	*Management Practice*	*Individual*	*Group*	*test of equality*
Finance	System of monitoring and control of financial goals in place	**0.83**	0.67	0.278
Finance	Frequency at which financial objectives and goals achieved	**0.80**	0.65	0.324
Finance	Existence of a Master Budget	0.72	**1.16**	0.014
Finance	Control and Monitoring of Master Budget	0.76	**1.02**	0.112
Finance	How deviations from Master Budget analyzed	**0.91**	**1.07**	0.380
Finance	Structure of Control and Monitoring Indicators (KPIs)	**0.94**	**0.96**	0.895
Finance	Agenda of Financial Management Meetings	**1.05**	**1.05**	1.000
HR	Strategic objectives leverage people's and team's talent	**0.83**	0.63	0.243
HR	Human talent development plans linked to corporate strategy	**0.81**	**0.90**	0.577
HR	Strategic plan defined, that includes clear goals for human talent	**0.95**	**0.91**	0.816
Marketing	Implementation of analysis of marketing trends	0.49	**0.87**	0.049
Marketing	Implementation of analysis of marketing risks	0.63	**0.90**	0.177
Marketing	Alignment of marketing and sales plan with business strategy	0.66	**0.82**	0.386
Marketing	Monitoring of sales behavior and trends	0.72	**0.90**	0.383
Production	Implementation of strategic goals between plant manager and supervisor	0.62	**0.97**	0.016
Production	Monthly monitoring of strategic goals between plant manager and supervisor	0.69	**0.89**	0.242
Production	Strategic goals and roles clear to each worker	0.67	**0.90**	0.067
Production	Each worker has improvement goals	0.56	**0.89**	0.030
Production	Bottlenecks are identified and managed	0.51	**0.84**	0.024
Production	Monthly measurement of plant KPIs	**0.82**	**0.86**	0.824
Production	Weekly or bi-weekly management of KPIs	**0.85**	0.65	0.281
Production	Improvement programs for KPIs developed	**0.93**	**0.99**	0.756
Production	Culture of visual management with graphs of machine performance	**0.81**	0.51	0.176
Production	Supervisors and workers manage improvement plans for quality anomalies	**0.80**	**0.94**	0.488

Notes:
Coefficients are post-intervention treatment effects for impact on individual management practices.

Associations between different measures of management and over time

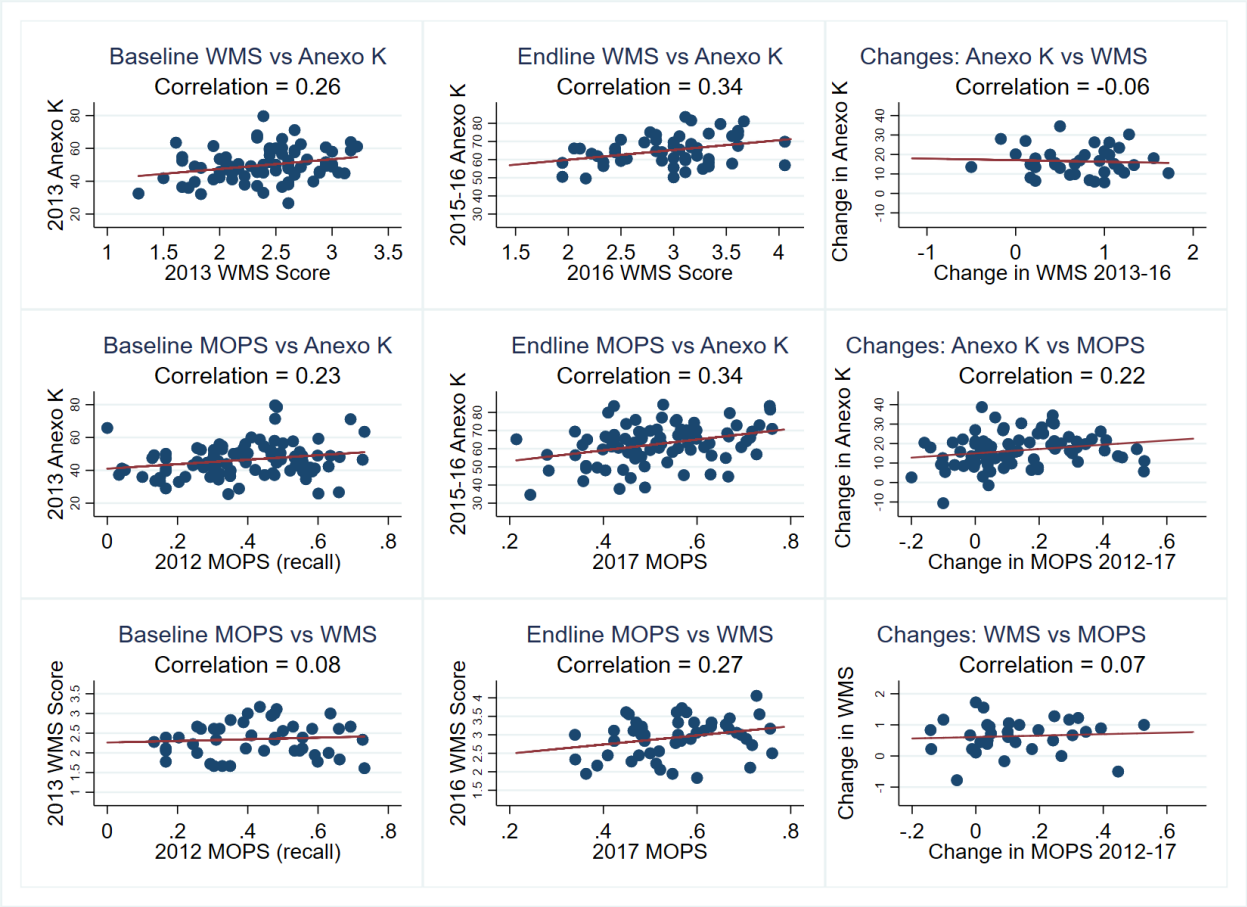
The WMS and MOPs are collected in a much less in-depth way than the Anexo K, and measure different aspects of management. Table A8.1 looks at the baseline correlations between different measures. At baseline, the Anexo K management score has a correlation of 0.26 with the WMS management score, and 0.23 with the MOPS score. By way of comparison, the 38 management practices in Bloom et al. (2013) had a 0.40 correlation with the WMS score. The Anexo K is most highly correlated with the monitoring component of the WMS (correlation of 0.44). When we examine the five areas of the Anexo K, the finance, logistics and production scores are more highly correlated with the WMS than the HR and marketing scores. Recall the WMS does not measure marketing practices, and there is a difference in emphasis in how the two focus on human resource practices. The WMS is more focused on how good and bad performers are hired and rewarded, whereas the Anexo K has more of an emphasis on organizational culture and links to overall business strategy. Notably, while the MOPS and WMS are intended to measure similar concepts, the correlation between the 2012 (recalled) MOPs management score and the WMS is only 0.08, suggesting substantial noise in this measurement.

Table A8.1: Correlations between baseline Management Measures

	WMS Overall	WMS Operations	WMS Monitoring	WMS Targets	WMS People	**MOPS Overall**
Anexo K Overall Score	**0.26**	0.16	0.44	0.04	0.11	**0.23**
Finance Score	0.28	0.22	0.46	0.07	0.07	0.15
HR Score	0.14	0.09	0.33	-0.08	0.03	0.17
Logistics Score	0.23	0.12	0.32	0.07	0.13	0.31
Marketing Score	0.09	0.03	0.12	0.02	0.06	0.10
Production Score	0.26	0.14	0.40	0.07	0.13	0.17
MOPS Overall	**0.08**	0.00	0.04	0.07	0.10	1.00

Figure A8.3 plots the cross-sectional and panel associations between measures. We see that the endline Anexo K has a cross-sectional correlation of 0.34 at endline with both the WMS and MOPS, and that the WMS and MOPS at endline still only have a correlation of 0.27. More starkly, there is no relationship between the WMS and Anexo K in the panel: firms which improve the most according to the Anexo K are unrelated to those which improve the most according to the WMS. This is also true of the association between changes in the MOPs and changes in the WMS. Recall that the WMS is done double-blind by phone, with enumerators scoring firms on a five-point scale. While there is signal in the responses, this also entails a lot of noise. Bloom et al. (2019) report that the test-retest correlation when two different people from within a plant answered the same questions within a few weeks of one another is only 0.51. This makes the 0.26 correlation at baseline between the Anexo K and WMS appear not so bad, especially given they are different survey instruments and were carried out months apart. In our case, there is an added factor of the baseline WMS being done by the LSE team, while the endline was collected by Innovations for Poverty Action (after training from the LSE team). As such, we should expect much of the change over time in the WMS to reflect measurement error, which can make it difficult to detect treatment effects.

Figure A8.3: Cross-sectional and panel correlations between management measures



Notes: first column shows cross-sectional correlations pre-treatment, second column shows cross-sectional correlations post-intervention for last measurement obtained by each method, and third column shows correlation of change in management (pre-post) according to each measure.

To investigate which of the three management measures is most strongly correlated with business outcomes of interest, we regress baseline log employment and labor productivity on each management measure separately, and then on all three together. The results are shown in Table A8.2. The Anexo K score is strongly associated with both log employment and labor productivity at baseline (both significant at the 1% level), while the WMS and MOPS have weaker associations. When all three measures are included together, the Anexo K measure remains statistically significant, while neither other measure is significant. This suggests the Anexo K measure has a stronger signal for business outcomes than these two alternatives.

Table A8.2: Baseline Association of Outcomes with Management

	Log Employment				Labor Productivity			
Anexo K Score	0.036***			0.017***	0.670***			0.877***
	(0.006)			(0.006)	(0.140)			(0.186)
WMS Management Score		0.250*		0.086		4.914		-0.652
		(0.134)		(0.153)		(4.070)		(5.310)
MOPS Management Score			0.869*	-0.554			8.994	-2.894
			(0.465)	(0.459)			(8.650)	(12.164)
Sample Size	159	77	95	46	159	77	95	46
R-squared	0.19	0.05	0.03	0.14	0.14	0.01	0.01	0.25

Notes:
Anexo K management practices are 141 management practices divided into five sub-areas.
WMS is World Management Survey, taken for subsample of firms in 2013. MOPS is Management and Organizational Practices Survey, and was conducted in 2017, with recall of practices 5 years earlier used to obtain baseline measure.
Robust standard errors in parentheses, *, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

Appendix 8: Impacts on World Management Survey and MOPS management measures

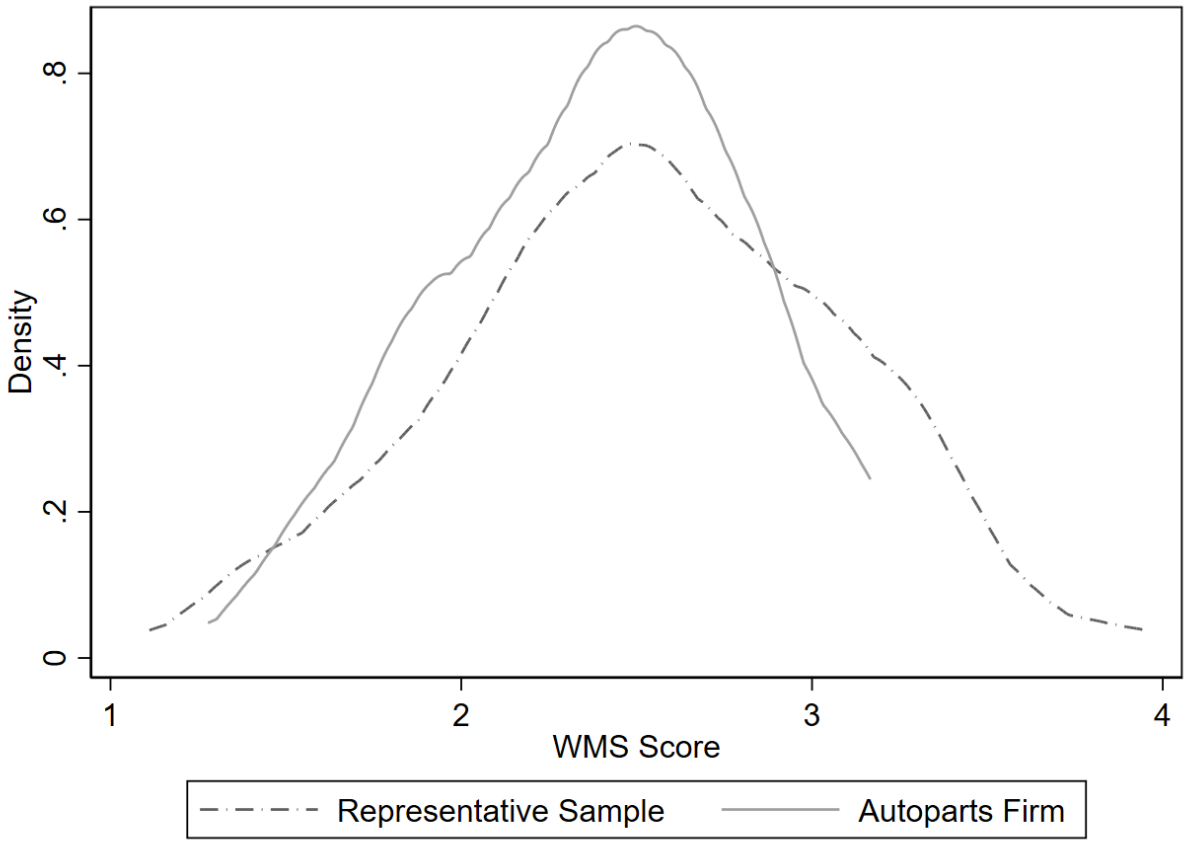
WMS 2013 Data Collection

We commissioned the London School of Economics (LSE) team responsible for the Bloom and Van Reenen (2007) World Management Surveys (WMS) to apply their methodology to a random sample of 180 firms representation of the Colombian manufacturing sector, as well as to a sub-sample of 77 firms in our sample, focusing on firms with 40 or more employees (Table A6.1).

Interviews were done by phone with a manager with thorough knowledge of the production process, typically the plant manager or production manager. The WMS interview is structured as a guided discussion, and is designed to be answered by a manager with thorough knowledge of the production process, typically the production or plant manager. Such discussion lasts between one hour and one hour and a half, and covers the 18 questions related to operations, monitoring, targeting, and people management. The interviewer guides the interviewee by means of open questions, letting him/her speak freely but making sure to have the necessary objective information to score each of the 18 topics using the provided scoring grid. Each of the 18 topics receives a score between 1 (no modern practice is implemented) and 5 (best practice).

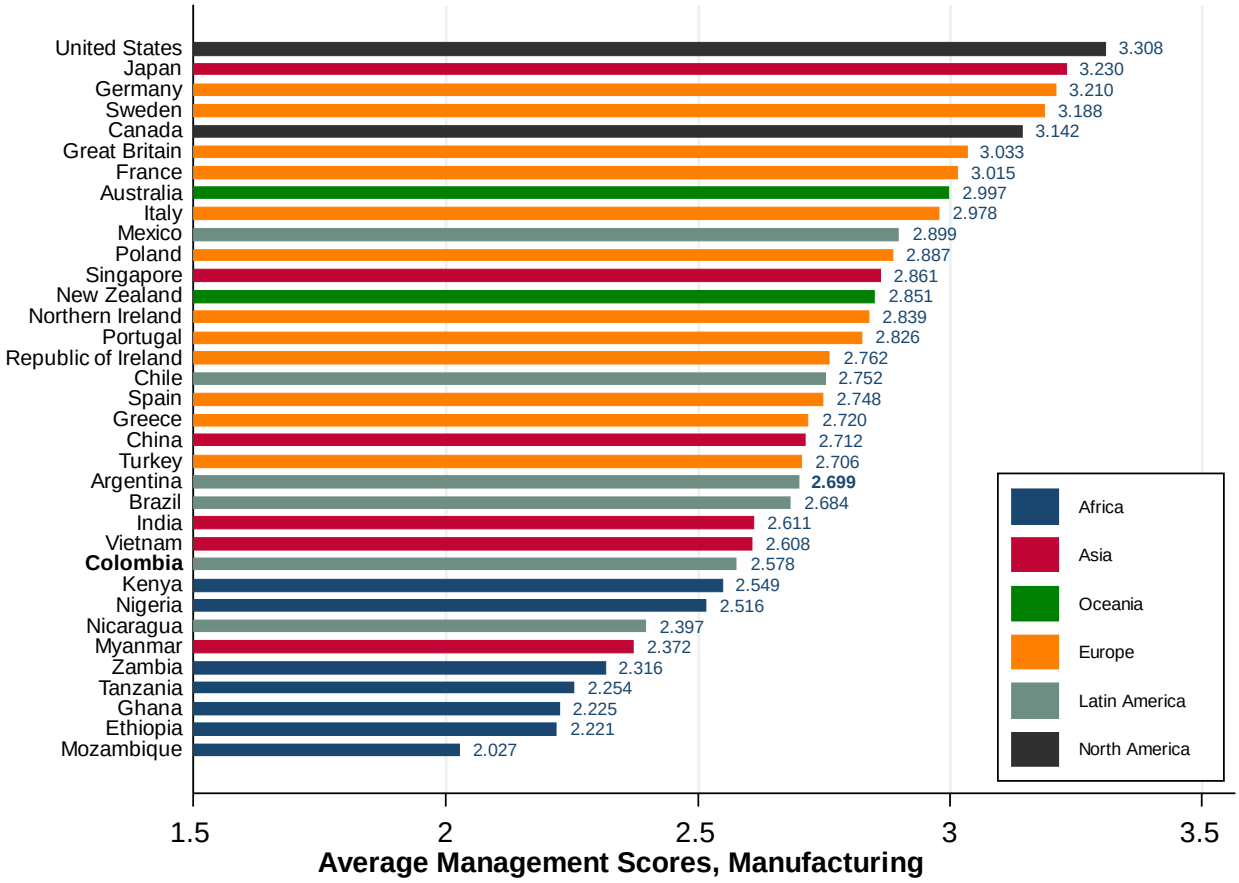
A first use of this survey was to be able to compare the management practices of the auto parts sector in our sample to that of Colombian manufacturing as a whole. Figure A8.1 shows that the distribution of management practices in our firms is similar to that of all SME manufacturing firms in Colombia. A second purpose was to enable comparison of Colombia to the rest of the world. Figure A8.2 shows Colombia's average management practices score of 2.54 are poorly managed by global standards, but typical for many developing countries, just below that of India and just above Kenya. The mean management practices score for the auto parts firms of 2.38 is similar.

Figure A8.1: Comparison of WMS Management Practices Distribution of our Auto Parts firms to a Representative Sample of the Colombian Manufacturing Sector



Source: WMS surveys conducted of 180 Colombian manufacturing firms and 77 auto parts firms conducted by the LSE WMS team in 2013.

Figure A8.2: Comparison of Colombian World Management Survey Management Score to Other Countries



Source: World Management Surveys.

WMS 2016 Data Collection

In September 2016, we asked *Innovations for Poverty Action* (IPA) to conduct a second round of the *World Management Survey (WMS)*. The LSE provided support in training the four analysts that conducted the interviews, the two supervisors and the research associate responsible for the survey. All material was provided by the LSE and the training took place in October 2016.

Since the WMS is designed for larger firms, we chose as a sample frame the 109 firms in our sample that had had at least 25 employees at baseline. This consisted of 37 control, 41 group treatment, and 31 individual treatment firms. Out of these 109 firms, we were able to collect data on 70 firms (20 control, 31 group, 19 individual), of which 50 firms had also been interviewed in 2013 (14 control, 22 group, 14 individual). This response rate of 64% is double the standard WMS response rate, reflecting the pre-existing contacts with these firms through the project. Of those companies not interviewed, 3 had closed down, and the remainder either refused, or repeatedly rescheduled and could not be interviewed.

Management and Organizational Practices Survey (MOPS)

Our final measure of management practices comes from a 16-question survey given to firm owners in 2017, derived from the Management and Organizational Practices Survey (MOPS). This survey was created by the U.S. Census bureau, and was designed to enable basic management practices to be measured in a self-administered survey format. The survey asks questions related to monitoring, targeting, and incentives, and is intended to measure similar concepts to the WMS (Bloom et al, 2019). It was carried out by Innovations for Poverty Action during in-person visits to the firms, and firms were also asked to recall what these practices were five years earlier (in 2012). Table A8.1 shows this data were able to be collected for 95 firms.

Treatment Effects on WMS and MOPS measures of management

Table A8.3 reports the estimated treatment impacts on the WMS and MOPS measures. Since these data are only available for a subset of our firms, we report several different specifications. In Panel A, we use all 70 firms for which follow-up WMS data are available (or the 95 firms with MOPS data for the last column). We do not control for randomization triplet fixed effects given that this would result in relatively few triplets being included. Instead, panel A includes no other controls, while Panel B controls linearly for key baseline variables used in the randomization (region, size, employment, labor productivity, and baseline Anexo K). Panels C through E then use the set of 50 firms for which both baseline and endline WMS data are available.

In panels A and B, we find very small and statistically insignificant impacts of either treatment on any of the WMS or MOPS management measures. Restricting to the sample for which we also have baseline data in panels C, D and E results in larger point estimates for the WMS, but the impacts are still far from statistically significant.

Our results show that both treatments resulted in significant increases in the Anexo K measure of management practices, and in each of its five subcomponents. This raises the question of why we do not see such a change in the WMS and MOPS? A first potential explanation is that the WMS and MOPS are only available for subsamples of the data, so that the difference in results could stem from sample composition and sample size. To investigate this hypothesis, Table A8.4 re-estimates the management treatment effect regressions for common sub-samples. The first column repeats our estimated impact on the Anexo K measure for the balanced panel. Columns 2 and 3 then consider the 52 firms for which we have both the 2016 WMS and Anexo K measured during and after the intervention. We continue to see a statistically significant impact of the individual treatment on the Anexo K measure using this sub-sample both during and post-intervention, and a significant impact of the group treatment during the intervention, with the magnitude of the estimated effect only falling in a substantive way for the group treatment post-intervention, although with a wide confidence interval. In contrast, there is no significant impact on the WMS using this same sample. The foot of the table converts the estimated treatment effects into confidence intervals expressed in terms of standard deviation changes in the respective management practice. We see that not only are the WMS treatment effects statistically insignificant while those for the Anexo K outcome are statistically significant, but the 95 percent confidence interval for the effect of the individual treatment effect does not even overlap for the two outcomes. This suggests that the lack of impact on the WMS is not simply a matter of the

sample composition or statistical power. Likewise, when we restrict to the same sample as the MOPS in columns 4 and 5, we find significant treatment impacts on the Anexo K, and no significant impact on the MOPS, although in this case the confidence intervals do overlap.

Table A8.3: Impact on Other Measures of Management Practices

	WMS Overall	WMS Operations	WMS Monitoring	WMS Targets	WMS People	MOPS Score
All firms interviewed in 2016						
Panel A: No controls						
Individual Treatment	0.040	0.100	0.152	-0.045	-0.003	-0.008
	(0.169)	(0.345)	(0.225)	(0.238)	(0.156)	(0.034)
Group Treatment	0.075	0.035	0.152	0.041	0.053	0.013
	(0.170)	(0.298)	(0.209)	(0.230)	(0.153)	(0.031)
Panel B: Baseline Controls						
Individual Treatment	-0.000	-0.030	0.095	-0.076	-0.007	-0.005
	(0.166)	(0.307)	(0.235)	(0.243)	(0.152)	(0.032)
Group Treatment	0.061	0.009	0.094	0.094	0.025	0.018
	(0.166)	(0.276)	(0.210)	(0.231)	(0.162)	(0.030)
Sample Size	70	70	70	70	70	95
Control Mean in 2016 of outcome	2.92	2.90	3.28	2.94	2.61	0.52
Control S.D. in 2016 of outcome	0.55	1.07	0.68	0.79	0.54	0.13
50 firms interviewed in WMS in 2013 & 2016						
Panel C: No Controls						
Individual Treatment	0.143	0.321	0.314	-0.086	0.131	0.010
	(0.218)	(0.423)	(0.256)	(0.311)	(0.199)	(0.051)
Group Treatment	0.283	0.357	0.312	0.225	0.284	0.064
	(0.216)	(0.363)	(0.254)	(0.293)	(0.183)	(0.045)
Panel D: Baseline Controls						
Individual Treatment	0.029	0.123	0.153	-0.188	0.074	-0.011
	(0.204)	(0.388)	(0.257)	(0.304)	(0.197)	(0.055)
Group Treatment	0.242	0.238	0.210	0.276	0.241	0.066
	(0.203)	(0.350)	(0.266)	(0.286)	(0.175)	(0.049)
Panel E: Baseline Controls + Ancova						
Individual Treatment	0.072	0.233	0.168	-0.160	0.133	-0.009
	(0.199)	(0.394)	(0.252)	(0.299)	(0.199)	(0.055)
Group Treatment	0.267	0.335	0.232	0.296	0.214	0.068
	(0.214)	(0.372)	(0.276)	(0.302)	(0.163)	(0.048)
Sample Size	50	50	50	50	50	46
Control Mean in 2016 of outcome	2.88	2.89	3.24	2.96	2.51	0.53
Control S.D. in 2016 of outcome	0.65	1.13	0.76	0.90	0.56	0.14

Notes:
Each panel represents treatment impacts from a separate regression.
70 of the 159 firms were given the WMS survey in 2016, of which 50 had also received this survey in 2013.
Panels A and C regress outcomes on treatment dummies only. Panels B and D add controls for
dummies for the Cundinamarca and Valle regions, a dummy for having 10 to 50 workers at baseline, the number of employees in 2013, labor productivity in 2013, and the 2013 Anexo K management practice score.
Panel E also controls for the baseline value of the outcome measure.
Robust standard errors in parentheses. *, **, and *** indicate significance at the 10, 5, and 1 percent levels.

Table A8.4: Impact on Anexo K on Same Samples as WMS and MOPS

	Balanced Panel	WMS Sample		MOPS Sample	
	Anexo K	Anexo K	WMS	Anexo K	MOPS
Individual Treatment*During Intervention	9.413***	8.350***		9.669***	
	(1.760)	(2.229)		(1.879)	
Individual Treatment*Post Intervention	9.309***	8.325***	-0.210	9.657***	0.017
	(1.821)	(2.368)	(0.176)	(1.856)	(0.036)
Group Treatment*During Intervention	11.384***	7.602**		11.143***	
	(2.202)	(3.164)		(2.438)	
Group Treatment*Post Intervention	8.155***	3.911	-0.132	7.549***	0.040
	(2.124)	(3.091)	(0.174)	(2.318)	(0.034)
Sample Size	202	104	52	172	86
Control Mean	55.98	60.1	2.93	57.44	0.49
Control SD	10.79	6.98	0.41	10.23	0.12
Implied 95% confidence intervals in S.D.					
Individual Treatment*Post Intervention	[0.53,1.19]	[0.53,1.86]	[-1.35,0.33]	[0.59,1.30]	[-0.45,0.73]
Group Treatment* Post Intervention	[0.37,1.14]	[-0.31,1.42]	[-1.15,0.51]	[0.29,1.18]	[-0.22, 0.89]

Notes:
Column 1 is for the 101 firms for which Anexo K management practices are measured both during and post intervention.
Columns 2 and 3 restrict to the subset of 52 firms that also had the WMS measured in 2016,
Columns 4 and 5 restrict to the subset of 86 firms that also had the MOPS measured in 2017.
Regressions control for baseline (December 2013) Anexo K mean, time fixed effects, and controls for region
baseline labor productivity, baseline number of employees, and for being a small firm at baseline.
Robust standard errors in parentheses, clustered at the firm level.
*, **, *** denote significance at the 10, 5, 1 percent levels respectively.

A more compelling explanation for the lack of impact on the WMS is due to this measure not being as able to pick up the types of changes in management practices that come from this intervention. A first reason for this is just the general noise in the measure, as discussed above. This noise means that much of the change in the WMS over time may reflect measurement error, making it difficult to detect treatment effects. But a second reason is that the WMS measures practices at a more general level than the level of specificity at which interventions are focused. Evidence in support of the idea that the WMS is not able to pick up the specific changes in practices that these consulting type interventions bring about comes from the India experiment that initially motivated this work. Bloom et al. (2013) report that their treatment plants increased their use of the 38 specific management practices they measure by 37.8 percentage points, significantly larger than the change for the control firms. They asked Accenture to also apply the WMS survey instrument to these firms during this post-intervention measurement phase. However, Accenture did not receive the LSE training on applying this survey instrument, and appear to have graded firms more harshly, with a mean WMS score of 1.45, compared to a baseline mean of 2.69 when conducted by the LSE team. Despite the large change in management practices observed in the 38 management practices used in Bloom et al. (2013), there is no significant difference in the follow-up WMS scores in this case (mean of 1.43 for the treated firms, 1.49 for the control firms, p-value = 0.693). So, as with our Colombian case, if one were to rely on the WMS to measure whether changes in management had occurred, the conclusion would have been that the Indian interventions had no significant effect on management.

Appendix 9: Comparison of PILA and Firm Employment Data and Changes in Composition of Firm Employment

The PILA is the platform through which firms pay social security data for their employees. We requested that government ministries with access to this data attempt to match our firms. This was done three times. First, the department of statistics (DANE) matched to the firm data between January 2014 and June 2016. Secondly, the Ministry of Health matched our firms to their database, covering the period January 2011 through February 2017, and then later re-matched for our firms from January 2012 through December 2018. Matching firms was not trivial, with firms' names not always given, the identification number of the company changing if the economic activity changes or some other features change, and at times the same firm being listed under the name of the owner versus the firm. The last attempt was the most successful and comprehensive, and our PILA series uses the second Ministry of Health extract as a base, correcting a small number of matching errors with data from the previous attempts.

Figure A9.1 shows a scatterplot of the employment reported in the PILA and the employment taken from the firm's records for the set of 7,010 year-month-firm observations between January 2013 and December 2017 for which we have data from both sources. The correlation is 0.94 over the full period, and the mass of points lie close to the 45-degree line. However, we do see a few points which have lower levels of employment reported in the PILA than in firm records. These likely reflect informal employment.

Figure A9.1: Employment Reported in PILA vs Employment Reported by Firms

In addition to data at the firm level, anonymized person-level data enable us to track inflows and outflows of workers from these firms, and to examine the gender and age composition of the workforce, as well as the monthly salaries paid workers. Column 1 of Table A9.1 looks at the proportion of workers that were working in firms in January 2013 who remained in the firm five years later, at the end of December 2017. In the control group, only 41 percent of workers are retained this length of time. The point estimate suggests a 10.7 percentage point increase in this retention rate in the group treatment firms, which is significant at the 10 percent level. Columns 2 and 3 show that 74 percent of workers are male and the average worker age is 43, with neither treatment having large, nor statistically significant impacts on these worker characteristics. Finally, Column 4 examines the treatment impact on mean worker monthly wages. The group treatment results in a 36,526 COP (3%) point estimate increase, but this is not statistically significant.

Table A9.1: Impact on Employment Composition

	Five-Year Retention: Proportion of Jan 2013 workers remaining in firm in Dec 2017	Worker Characteristics		
		Proportion Male	Mean Age	Mean Monthly Salary (COP)
Individual Treatment*During Intervention		0.008	-0.052	-38509
		(0.009)	(0.362)	(29109)
Individual Treatment*Post Intervention	0.029	0.009	0.130	-37387
	(0.051)	(0.014)	(0.485)	(36636)
Group Treatment*During Intervention		-0.001	-0.041	-36
		(0.008)	(0.404)	(25805)
Group Treatment*Post Intervention	0.107*	-0.002	-0.272	36526
	(0.059)	(0.010)	(0.467)	(33202)
Sample Size (N*T)	135	8502	8502	8502
Sample Size (N)	135	146	146	146
P-value: Individual=Group During		0.472	0.985	0.339
P-value: Individual=Group Post	0.168	0.500	0.482	0.117
Control Mean	0.41	0.74	43.0	1088858
Control S.D.	0.20	0.14	4.97	427825

Notes:
Regressions use PILA data on formal employment, and are for sample of surviving firms.
Column 1 is a cross-sectional regression for firms with employment data in both Jan 2013 and Dec 2017.
Columns 2, 3 and 4 include firm and time fixed effects, and cluster standard errors at the firm level.
*, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

We also use this data in Table A9.2 to examine more closely the characteristics of the workers who moved across firms in our experiment. We compare the workers who work for only one firm in our sample during January 2013 to December 2017 with workers who work for two or more firms in the sample. Within the workers who switch firms, we also consider separately the subset who switched to a firm in a different treatment status. We see that male workers are more likely to switch firms. The workers switching are of similar age and earn similar salaries to those not switching. While we cannot identify the position in the firm, we expect managers to be older and

earn more than the average worker, and so this suggests that the flow of workers across firms is not disproportionately made of managers.

Table A9.2: Comparison of Characteristics of Workers who Switch Firms to those who don't

	Sample Size	Proportion Male	Mean Age	Mean Salary	Median Salary
Work for only one firm in sample	22,884	0.717	36.3	946625	720264
Work for two or more firms in sample	272	0.897	36.3	911527	737254
Work for firms in two different treatment groups	32	0.813	35.7	995477	720437
P-value: test that mean for switchers= non-switchers		0.000	0.925	0.581	
P-value: test that mean for treatment switchers=non-switchers		0.231	0.782	0.792	

Note: data on workers in the PILA who work for at least one treated firm in at least one month from Jan 2013-Dec 2017
Salary is monthly and is in Colombian pesos. Exchange rate in 2014-15 is 2372 COP = 1 USD.

Appendix 10: Time Since Treatment, and Our Survey Data on Sales

Table A10.1 examines how the impacts of treatment vary with time since treatment, and show that we cannot reject equality of treatment effects over time.

Table A10.2 uses the data on monthly sales that CNP and IPA collected directly from firms. Although firms were initially willing to share sales data with us, the long timespan of the project and survey fatigue led many firms to stop sharing data over time. From our direct survey efforts we have monthly sales data for some months post-baseline for 145 firms, but only have 99 of these with data for all 60 months between January 2013 and December 2017 (which is still a year earlier than the EAM data runs). We use firm fixed effects to account for time invariant characteristics of firms that may be correlated with sample attrition. Columns 1 and 2 of report the estimated impacts on the levels of monthly sales for the unbalanced and balanced panel respectively, and columns 3 and 4 report impacts using the inverse hyperbolic sine. The group treatment has positive treatment effects on monthly sales, of 63-71 million COP per month (USD $26,500-$29,900) in levels, or 9 to 10 percent in log terms, but these impacts are not statistically significant. The individual treatment effects are smaller in magnitude, and even negative in the level results, and also are not statistically significant.

Table A10.1: Time-Varying Impacts on Employment, Profits and Sales

	PILA Data Jan 2012-Dec 2018 Conditional		EAM Data Annual Data 2010-2018			
			Profits		Sales	
	Level	I.H.S.	Levels	Logs	Levels	Logs
Individual Treatment*During Intervention	-0.681	-0.063	23.15	-0.245	306.5	-0.100
	(3.037)	(0.043)	(305.2)	(0.218)	(557.2)	(0.095)
Individual Treatment* Year 1 Post	1.613	-0.004	-332.3	-0.275	333.5	0.004
	(3.948)	(0.050)	(415.1)	(0.267)	(830.4)	(0.130)
Individual Treatment* Year 2 Post	3.540	0.064	-124.5	-0.253	505.1	0.122
	(5.706)	(0.084)	(656.3)	(0.271)	(1002)	(0.104)
Individual Treatment* Year 3 Post	3.544	0.044	193.4	0.0462	1.010	0.125
	(6.121)	(0.101)	(612.2)	(0.183)	(1031)	(0.110)
Individual Treatment*Year 4 Post	3.503	0.014	319.1	-0.135	1091	-0.048
	(6.615)	(0.119)	(687.4)	(0.183)	(1204)	(0.141)
Group Treatment*During Intervention	5.649	0.127**	557.5**	-0.318**	1092**	0.229***
	(3.789)	(0.061)	(272.5)	(0.140)	(494.1)	(0.082)
Group Treatment* Year 1 Post	8.485*	0.202**	527.5	0.190	1903***	0.300***
	(4.775)	(0.086)	(335.0)	(0.179)	(676.9)	(0.097)
Group Treatment* Year 2 Post	7.428	0.196**	658.1	0.305*	1896***	0.276***
	(5.379)	(0.099)	(414.6)	(0.157)	(720.9)	(0.104)
Group Treatment* Year 3 Post	6.371	0.174*	633.9	0.260*	1793**	0.237**
	(5.898)	(0.103)	(439.3)	(0.153)	(822.6)	(0.110)
Sample Size (N*T)	11807	11807	1008	962	1008	1008
Number of Firms	147	147	120	118	120	120
P-value: Individual Year 1 = Year 2 = Year 3 = Year 4	0.905	0.394	0.471	0.305	0.515	0.062
P-value: Group Year 1 = Year 2 = Year 3	0.696	0.785	0.854	0.594	0.940	0.617
Control Mean in 2013	59.29	4.42	2541	14.85	5580	15.09
Control S.D. in 2013	51.95	0.89	2680	1.22	5439	0.992

Notes:

Fixed effects regressions with firm and time fixed effects. Standard errors clustered at the firm level are in parentheses.

*, **, *** denote significance at the 10, 5, and 1 percent levels respectively.

Level denotes level of outcome used; I.H.S. is inverse hyperbolic sine transformation.

PILA data are formal employment data from administrative records.

Conditional is for the group of surviving firms.

Table A10.2: Impacts on Monthly Firm Sales Using Our Survey Data

	Firm Survey Data Jan 2013-Dec 2017			
	Winsorized Levels		Inverse Hyperbolic Sine	
	(1)	(2)	(3)	(4)
Individual Treatment*During Intervention	-5	-22	0.054	-0.026
	(24)	(30)	(0.044)	(0.044)
Individual Treatment*Post Intervention	-21	-38	0.049	0.029
	(33)	(37)	(0.068)	(0.075)
Group Treatment*During Intervention	46	44	0.080	0.086
	(46)	(53)	(0.061)	(0.069)
Group Treatment*Post Intervention	67	63	0.103	0.091
	(44)	(48)	(0.084)	(0.093)
Balanced Panel	No	Yes	No	Yes
Sample Size (N*T)	7343	5940	7343	5940
Number of Firms	145	99	145	99
P-value: Individual=Group During	0.358	0.305	0.743	0.211
P-value: Individual=Group Post	0.117	0.099	0.519	0.486
Control Mean in 2017	388	407	5.994	6.033

Notes:
Fixed effects regressions with firm and time fixed effects. Standard errors clustered at the firm level in parentheses. *, **, *** denote significance at the 10, 5, and 1 percent levels respectively.
Monthly sales are measured in millions of real Colombian pesos, with levels winsorized at the 95th percentile.