DISCUSSION PAPER SERIES

DP14284

LEARNING MANAGEMENT THROUGH MATCHING: A FIELD EXPERIMENT USING MECHANISM DESIGN

Marcel Fafchamps, Girum Abebe, Michael Koelle and Simon Quinn

DEVELOPMENT ECONOMICS

ISSN 0265-8003

LEARNING MANAGEMENT THROUGH MATCHING: A FIELD EXPERIMENT USING MECHANISM DESIGN

Marcel Fafchamps, Girum Abebe, Michael Koelle and Simon Quinn

Discussion Paper DP14284
Published 06 January 2020
Submitted 19 December 2019

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

LEARNING MANAGEMENT THROUGH MATCHING: A FIELD EXPERIMENT USING MECHANISM DESIGN

Abstract

We place young professionals into established firms to shadow middle managers. Using random assignment into program participation, we find positive average effects on wage employment, but no average effect on the likelihood of self-employment. We match individuals to firms using a deferred-acceptance algorithm, and show how this allows us to identify heterogeneous treatment effects by firm and intern characteristics. We find striking heterogeneity in self-employment effects, and show that some assignment mechanisms can substantially outperform random matching in generating employment and income effects. These results demonstrate the potential for matching algorithms to improve the design of field experiments.

JEL Classification: N/A

Keywords: field experiments, Management Practices, Self-employment, causal inference, Propensity score

Marcel Fafchamps - fafchamp@stanford.edu
Stanford University and CEPR

Girum Abebe - girumabe@gmail.com
Policy Studies Institute, Addis Ababa

Michael Koelle - michael.koelle@economics.ox.ac.uk
University of Oxford

Simon Quinn - simon.quinn@economics.ox.ac.uk
University of Oxford and CEPR

Acknowledgements
We are grateful to Attila Ambrus, Nick Bloom, Jonathan Colmer, Matthew Elliott, Avi Feller, Ben Golub, Seema Jayachandran, Terence Johnson, Shengwu Li, David McKenzie, Eva Meyersson Milgrom, Benjamin Roth, Tom Schwantje, Ran Shorrer, Meredith Startz, Alex Teytelboym, Gabriel Ulyssea and Chris Woodruff for useful comments, and to audiences at Bocconi University, Bo˘gaziçi University, DFID, the EBRD, Harvard University, Stanford University and the University of Oxford. We thank Gezahegn Gebremedhin, Biruk Tekle and Alemayehu Woldu for outstanding research assistance. We thank the UK Department for International Development for research funding (through the Private Enterprise Development in Low-Income Countries program). Our pre-analysis plan is available at https://www.socialscienceregistry.org/trials/2776/history/49533.

Learning Management Through Matching: A Field Experiment Using Mechanism Design[*]

Girum Abebe[†], Marcel Fafchamps[‡], Michael Koelle[§] and Simon Quinn[¶]

August 2019

Abstract

We place young professionals into established firms to shadow middle managers. Using random assignment into program participation, we find positive average effects on wage employment, but no average effect on the likelihood of self-employment. We match individuals to firms using a deferred-acceptance algorithm, and show how this allows us to identify heterogeneous treatment effects by firm and intern characteristics. We find striking heterogeneity in self-employment effects, and show that some assignment mechanisms can substantially outperform random matching in generating employment and income effects. These results demonstrate the potential for matching algorithms to improve the design of field experiments.

KEYWORDS: field experiments, management practices, self-employment, causal inference, propensity score.

[*] We are grateful to Attila Ambrus, Nick Bloom, Jonathan Colmer, Matthew Elliott, Avi Feller, Ben Golub, Seema Jayachandran, Terence Johnson, Shengwu Li, David McKenzie, Eva Meyersson Milgrom, Benjamin Roth, Tom Schwantje, Ran Shorrer, Meredith Startz, Alex Teytelboym, Gabriel Ulyssea and Chris Woodruff for useful comments, and to audiences at Bocconi University, Boğaziçi University, DFID, the EBRD, Harvard University, Stanford University and the University of Oxford. We thank Gezahegn Gebremedhin, Biruk Tekle and Alemayehu Woldu for outstanding research assistance. We thank the UK Department for International Development for research funding (through the Private Enterprise Development in Low-Income Countries program). Our pre-analysis plan is available at https://www.socialscienceregistry.org/trials/2776/history/49533.

[†] **Policy Studies Institute, Addis Ababa**: girumabe@gmail.com

[‡] **Freeman Spogli Institute, Stanford University**: fafchamp@stanford.edu

[§] **Pembroke College and Centre for the Study of African Economies, University of Oxford**: michael.koelle@economics.ox.ac.uk

[¶] **Department of Economics, Centre for the Study of African Economies and St Antony's College, University of Oxford**: simon.quinn@economics.ox.ac.uk.

1 A novel matching experiment

Good management is both important and scarce in developing countries. Differences in structured management practices are associated with productivity differences in firms throughout the size distribution, and firms in the developing world tend to have the worst management practices (Bloom and Van Reenen, 2007; McKenzie and Woodruff, 2017). Management also matters at the individual level (Bandiera, Hansen, Prat, and Sadun, 2017): managers who are not effective leaders of their organizations may forego productive investments (Atkin, Chaudhry, Chaudry, Khandelwal, and Verhoogen, 2017) or fail to recruit the most productive workers (Abebe, Caria, and Ortiz-Opsina, 2018). Less is known about how future managers are made.

To become a successful manager, individuals must gain experience in many practical tasks. They must learn how to monitor performance against targets, how to organise relationships with clients, suppliers, and colleagues, and how to navigate the business environment. The skills required to set up and manage modern organizations are difficult to learn either in the classroom or from scratch — for example, by starting a microenterprise (Brooks, Donovan, and Johnson, 2018). Managing is typically learned on the job, by observing practical situations and the actions taken to address them (Ellison and Holden, 2013). Evidence from high- and middle-income countries suggests that many high-productivity enterprises are started by entrepreneurs who have spent several years in wage employment (Humphries, 2017) or are spun off by former employees of large firms (Muendler and Rauch, 2018; Klepper and Sleeper, 2005). In developing countries, the kind of large and efficient organizations that can incubate such skills often are few and far between.

We run a novel field experiment, designed to test the effects of exposing young professionals to experienced managers, to test for heterogeneity across different varieties of treatment, and to evaluate the importance of matching in treatment assignment. We create a new institution to match and place highly educated young professionals inside established medium and large firms where they spend a one-month 'management experience'. Management placement is a potentially valuable way of learning and diffusing managerial capital, particularly in a developing country context. We act as market maker by identifying a suitable population of firms and a pool of early-stage young professionals and by matching the professionals with internships in firms. The field experiment takes place in Ethiopia, a fast-growing economy with few established private-sector firms. The experiment gives each randomly selected participant the opportunity to shadow a middle manager in their daily work, in order to see and experience first-hand the managerial realities of medium and large scale organizations. Host firms are representative of businesses in Addis Ababa and surroundings, and operate in different sectors, at different scale, and with different management practices.

We use two different assignment mechanisms in our experiment — each of which facilitates a different kind of causal estimation. First, we use *random assignment to select which participants enter the program* and which are assigned to the control group. This allows us to the estimate average treatment effect of being invited to the program. We evaluate the impact on young professionals' labor market outcomes, and on their management skills, attitudes, and practices. On average, we find that offering a management placement to highly educated and motivated young Ethiopians increases their success in obtaining a good permanent wage job, and increases their hours and earnings in wage employment. It has no average effect on their propensity to run a firm on their own, and no significant effect on the profits of firms run by this group. Respondents — whether self-employed or not — report an increased confidence in their management abilities along several dimensions that are consistent with the tasks that the average participant would have completed.

Second, among those participants entering the program, we use a *firm-proposing deferred-acceptance algorithm to match individuals with host firms*. The controlled structure of this matching algorithm allows us to develop a novel empirical strategy to characterise heterogeneous treatment effects by both firm and participant. Specifically, we exploit the fact that the matching algorithm is implemented in small batches to which subjects are invited from a randomly ordered list. We estimate a generative Bayesian classification model to characterize the determinants of individuals' ranking of potential host firms, and the rankings of interns by firms. The results of this model are then used as a basis for integrating over many plausible counterfactual assignments, exploiting the 'equal treatment of equals' property of our controlled matching to simulate match propensities. We build on recent empirical advances from the mechanism design literature on school choice (Abdulkadiroğlu, Angrist, Narita, and Pathak, 2017) and we apply Rosenbaum and Rubin (1983)'s propensity score theorem to identify the causal effect of being assigned to a specific type of host firm. The method is applicable to any setting where assignment to treatment is achieved by matching participants: (i) by a controlled and fully replicable mechanism; (ii) undertaken in small batches that are randomly populated; and (iii) for which the distribution of rankings and choice lists across batches is known or can be estimated. This covers a broad array of interventions, such as consulting and training (Brooks, Donovan, and Johnson, 2018; Bruhn, Karlan, and Schoar, 2018), education (Banerjee, Cole, Duflo, and Linden, 2007), and matching in networks (Breza and Chandrasekhar, 2019).

To serve our research objective while ilustrating the usefulness of the experimental design, we test the causal effect of assignment to a host firm with good management practices. We do this by recovering the marginal treatment effect (MTE), defined in our context as the differential effect of being assigned to a high-management firm as a function of the propensity score (Heckman, 2005; Carneiro, Heckman, and Vytlacil, 2011). We find striking heterogeneity in self-employment effects but almost no heterogeneity on wage employment. Specifically, we find a positive effect on self-employment of being matched to a high-management firm. Furthermore this effect is higher and significantly different from zero for participants who have a high propensity of being assigned to a high-management firm – that is, those who rank high-management firms highly and are highly ranked by them. For these subjects, assignment to a high-management firm increases the probability of self-employment by 3-4 percentage points relative to assignment to a low-management firm.

These results begs an important follow-up question: given that treatment effects are heterogeneous based on the identity of both host firm and individual participant, *is there any value in using a firm-proposing deferred acceptance algorithm to match interns and firms?* or *would an alternative algorithm be more effective?* To answer this, we extend the propensity score method to simulate counter-factual assignment mechanisms. This is achieved by comparing the propensity scores obtained under our design with those obtained under alternative mechanisms. We find that firm-proposing deferred acceptance substantially outperforms random matching in generating employment income effects. But we also find that higher employment income effects would have been generated by using an algorithm based on intern-proposing deferred acceptance or intern-proposing random serial dictatorship.

These results have important implications for the design of field experiments. Treatment effects can be purposefully affected by the mechanism by which treatment type is assigned within the treatment group (in our case, the assignment of interns to firms). We demonstrate how researchers can evaluate the payoffs from a class of counterfactual assignment mechanisms without resorting to additional experimentation. In our particular application, the results suggest that an assignment mechanism that respects the preferences of the treated can improve their welfare. This is consistent with the theoretical findings in Narita (2018).

Our work relates to two distinct bodies of literature. First, the results speak to a literature in organizational and managerial economics about management heterogeneity (Bertrand and Schoar, 2003; Ellison and Holden, 2013; Bandiera, Guiso, Prat, and Sadun, 2015; Bandiera, Hansen, Prat, and Sadun, 2017; Gosnell, List, and Metcalfe, 2019). This literature distinguishes two prominent sources of heterogeneity in management: differences in the traits and characteristics of managers; and differences in management style at the interplay between manager and organization. To our knowledge, there has been almost no empirical research on this theme in developing economies.[1] Understanding this process is potentially important for firm productivity, and ultimately for the aggregate performance of less developed economies, for at least two reasons. On the one hand, organizational management practices in developing countries are on average less developed and more dispersed (Bloom, Lemos, Sadun, Scur, and Van Reenen, 2014). On the other hand, the productivity dispersion across firms is more pronounced in less developed countries (Hsieh and Klenow, 2009).

Second, from a methodological perspective, our paper adds to a growing body of research that applies theoretical insights and approaches from mechanism design to field experiments, in particular in developing countries (Jayachandran, de Laat, Lambin, Stanton, Audy, and Thomas, 2017; Rigol, Hussam, and Roth, 2018; Narita, 2018). While most of this line of work employs mechanism design to elicit truthful reporting or to incentivize behavior, our approach utilizes mechanism design for encouraging program take up and most importantly, in the service of causal inference. In this respect, our work shares similarities with the empirical literature on school choice mechanisms since they serve similar purposes (Abdulkadiroğlu, Angrist, Narita, and Pathak, 2017). Our results also suggest that assignment by mechanism can increase program effectiveness. In this respect our work is related to the literature on using mechanism design to improve central allocation outcomes (Trapp, Teytelboym, Martinello, Annderson, and Ahani, 2018).

Our paper proceeds as follows. In Section 2 we describe the progam, experimental design and context, data, and implementation. We present average effects of the program from treatment-control comparison in Section 3. We develop an empirical framework for identifying causal effects using our controlled assignment mechanism in Section 4, and we apply this framework to estimate differential effects by treatment type and marginal treatment effects. In Section 5 we evaluate and compare counterfactual assignment mechanisms. We conclude in Section 6.

[1] A limited quantitative literature in finance studies these kinds of questions in developed countries — for example, see Benmelech and Frydman (2015) on military CEOs, Kaplan, Klebanov, and Sorensen (2012) and Malmendier, Tate, and Yan (2011) more generally on personality traits and skills, and Guiso, Sapienza, and Zingales (2015) on corporate culture.

2 The experiment

2.1 Experimental design

We conducted a two-sided field experiment with firms and young professionals in Addis Ababa, the capital city of Ethiopia, and in the nearby towns of Adama and Bishoftu. The experiment involved a novel programme that we designed: placing young professionals in established firms where they would shadow a manager in his or her daily activities, for a month. This modality of the placement program resembles internships — which are known in this context for certain professional careers — and we were careful to emphasize to all parties involved that the purpose was to learn about the day-to-day activities of a manager.

The fieldwork was carried out in 2016 and 2017 by staff at the Ethiopian Development Research Institute. Both sides — young professionals and firms — were randomised to participate in the program. Young professionals and firms in the treatment groups were paired with an algorithmic mechanism that we describe below. We evaluate the impact of this program on young professionals' labor market trajectories and outcomes, their skills, knowledge and attitudes about management and entrepreneurship, as well as potential channels that that link treatment to these outcomes of interest. Given the two-sided randomisation, we can also detect any potential impacts on firms.[2]

Our experiment was designed for a high-skilled population, having an expressed interest in management and entrepreneurship. We recruited with this aim in mind: we limited eligibility to young Ethiopians (aged 18 to 30, inclusive), having a minimum of technical/vocational, college or university qualifications (the 'young professionals'). We advertized using a combination of social media, college campus visits and postings on city 'job boards', using a headline message designed to attract aspiring managers and entrepreneurs. We approached firms in two ways. First, we recruited among firms in Addis Ababa that had already been part of another study (Abebe, Caria, Fafchamps, Falco, Franklin, and Quinn, 2018). Second, we updated that firm listing with new data that had become available in the interim, as well as with data provided by the respective municipal autorities in Adama and Bishoftu. We drew new random samples, weighted by industry employment shares, and subject to a minimum size and turnover threshold. Our experimental sample consists of firms that completed a baseline survey and confirmed their interest to our enumerators. We

[2] Appendix A contains further details on recruitment of young professionals, benchmarking of professionals and firms, the randomization procedure for young professionals, the implementation of the DA algorithm, and a short summary of the program implementation based on a debriefing survey and administrative records we collected.

created gathered fields of these firms according to their date of completion of the baseline survey (subject to availability expressed by the firm) and randomized firms by computer within each gathered field.

2.2 Context and data

Addis Ababa is an excellent location to test this sort of program. The Ethiopian economy has been growing steadily over the last two decades, admittedly from a very low initial level. The country is landlocked and has a large population, creating a captive market for locally produced goods. With no colonial legacy but a recent history of socialist rule, the country has few established foreign investors and suffers from a shortage of large, well managed firms, although things are changing.

In line with the eligibility criteria we specified, the experimental sample are drawn from a highly educated sub-population (Table 1). Three out of four completed at least an undergraduate college degree; and the remainder possess a vocational or technical post-secondary qualification. The median schooling is 15 years. The most frequent degrees are in engineering and related STEM subjects, and in business studies. We encounter these individuals at a transition period in their lives. Half the sample graduated in the year before they entered our program, or in the same year. At the moment of their induction session, only 25% of participants are employed in some form of wage job, and another 7% run their own business. 80% have undertaken any job search activity in the last month, and all of those who search do so for a professional or managerial job. 30% have thought of starting a business, but only few went from having an idea to taking actual steps. When we conduct a follow-up interview a year later, almost 70% of the control group have found a wage job, and 13% are self-employed.[3] It is in this context of rapid entry into a professional career that our intervention takes place.

Host firms are medium to large establishments operating in a variety of sectors in the economy. Most firms are located in Addis Ababa, while around 15% are located in two mid-sized towns in central Ethiopia: Adama (80km from Addis Ababa) and Bishoftu (40km from Addis Ababa). The firms operate mainly in services (about 40%), manufacturing (about 25%), trade (about 20%) and other sectors. The median firm has 57 employees ($Q_1 = 22$, $Q_3 = 155$). The firms are nationally representative in their size distribution. Man-

[3] The job seekers seemed to have had realistic expectations about the wages they could eventually earn – see Appendix Figure A.5 which compares reservation wages to realized wages in the sample at baseline and a year later. Expectations about profits in self-employment seem to be placed too high for at least the bottom half of the distribution (see Appendix Figure A.6).

agement practices of these firms are towards the bottom of the cross-country distribution reported by Bloom, Schweiger, and Van Reenen (2012).[4] Our questionnaire embeds their survey instrument and hence management practice scores are directly comparable.

2.3 Implementation

Our randomisation yields balanced treatment and control samples. For young professionals, we applied pairwise stratified randomisation on education, age and gender. Our randomization is strongly balanced on all outcome variables, as we report in Appendix Table A.3. Firms were implicitly stratified on the date of their baseline survey, but not paired. We find that firms are balanced on most variables other than management scores. This is due to treated firms being more likely to monitor any production performance indicators; a survey question which nests several other indicators on management practices. We control for any baseline differences in the dependent variable through ANCOVA.

For logistical reasons, we conducted the program on a rolling basis, in 42 weekly batches with on average 40 young professionals (of whom about 20 would be assigned a placement) and 8 firms per group. We purposefully selected firms into batches according to their availability. This proved to be important to ensure the participation of firms. We invited applicants in a random order to an information session about the program.[5] We randomized participation in the program among the young professionals who turned up at each induction session using a physical randomization device. The Appendix provides further details. We refer to the young professionals that were assigned to participate as 'interns'.

Within each batch, we invited both the interns and the firms to rank the respective other group, based on a small number of variables that we picked. Firms were given a short (anomymous) CV of interns that included gender, age, education level, field of study, and higher education institution, and work experience (both in time and by industry). Appendix Figure A.7 shows the form we used for the CV. Interns were given the following information about the firm: name of the firm, sector, approximate location, and size category of the firm. We enforced completeness and transitivity of the rankings: all interns had to rank all firms within a batch (and vice versa) and ties were not allowed.

We then matched all interns and firms within the batch, using the deferred-

[4] We show this descriptively in Appendix Figure A.4.

[5] We accepted applications for several rounds, and randomized the order within each round. Our estimation strategy allows for the fact that the pool of applicants may have changed over the course of the experiment; it only requires participants to be drawn randomly across nearby batches.

acceptance (DA) stable matching algorithm first proposed by Gale and Shapley (1962). We describe our implementation of DA matching in the Appendix. In our algorithm, firms are the 'proposing' side. The rule of assignment by mechanism was strictly followed and monitored by our field staff. Firms and young professionals were notified of their respective match and in general, participants would start their placement in the week following their induction sessions, typically on the Monday. Generally, participants were accompanied and introduced to the firm on their first day of placement by one of our field staff.

The main result from assigning placements by DA matching is that most interns and firms end up with a preferred counterpart. In Figure 1, we display the relative rank of the counterpart that interns and firms were eventually matched with. Around 40% of interns and 45% of firms end up with a counterpart in their top 20%. In most batches, this corresponds to interns being assigned to their top 1 or 2 firm. Firms have slightly more preferred counterparts, which would be expected since they are the proposing side in our DA algorithm (Gale and Shapley, 1962; Roth and Sotomayor, 1990). Overall, more than 70% of interns and firms end up matched with a counterpart in their top 40%. As a further descriptive statistic on this matching context, we implement the algorithm of McVitie and Wilson (1971), to calculate the size of the core of stable matches for each of the batches. We find that the cores are generally reasonably small: the median batch has just two stable matches (that is, the firm-proposing deferred acceptance solution and the intern-proposing deferred acceptance solution), and the range is from one to six. Appendix Table A.5 lists the details for each batch.

We can define take-up of our program by young professionals in several ways, shown in Appendix Table A.6. Of all the 829 young professionals assigned to treatment, 788 (95%) completed the process of being matched to a firm and 588 (71%) completed at least one day at the firm. From qualitative evidence and an exit survey with treated young professionals, the most common reason for not completing the placement seemed to have been: holding or finding a job with no possibility to take leave of absence, the location of the firm, and personal reasons such as family issues.

3 Average treatment effects

For analysing and reporting average effects of the randomized assignment to an internship placement, we follow a detailed pre-analysis plan.[6] We pre-specified the estimation equations, families of outcome variables of interest, the definition of these variables, subgroup

[6] This is available at www.socialscienceregistry.org/trials/2776.

analysis to be carried out, and our approach to multi-hypothesis testing and attrition. We note the experimental treatment effects that we estimate here are combining both the effect of being offered a placement *and* the particular assignment mechanism that we used; we unbundle these two elements in Section 5 below.

For some individual respondent i, denote T_i as a dummy for whether i was randomized into placement. Randomization was carried out in stratified pairs; we index the corresponding pair dummies by p. We observe each individual at baseline (which we denote as $t = 0$), at a six-month follow-up (which we denote $t = 1$) and at a 12-month follow-up ($t = 2$). Our preferred estimating equation is ANCOVA with pairwise dummies; that is, for individual i in pair p at time $t > 0$, we estimate:[7]

$$y_{ipt} = \beta_1 \cdot T_i + \beta_2 \cdot y_{ip0} + \delta_p + \varepsilon_{ipt}. \tag{1}$$

Our coefficient of interest is β_1, which we interpret as the *intent-to-treat* (ITT) estimate. For each hypothesis test on the coefficient of interest, we report the usual p-value from a Wald test; and the report False Discovery Rate q-values, taken across the coefficients of interest within an outcome family (Benjamini, Krieger, and Yekutieli, 2006). We winsorize all continuous outcome variables (within each survey wave) at the 95th percentile.

3.1 Occupation and earnings

Our first main area of interest is the impact of the management placement on young professionals' labor market outcomes — in particular, occupation and earnings. The placement could potentially have eased managerial capital constraints in entrepreneurially minded young professionals and help them start a business. Alternatively, it could have helped them in their search for a wage job, especially for professional and managerial positions. We estimate the effect of the program on occupation choices, hours and earnings in both self-employment and wage-employment. Table 2 presents the results. Below the point estimates, we report standard errors in parentheses, p-values in square brackets and q-values in curly brackets.

Our estimates imply that, on average, the program had no effect on self-employment. For occupation, this zero effect is precisely estimated both at the extensive and inten-

[7] Neither our sampling process nor the randomization was clustered; therefore, following the recent guidance of Abadie, Athey, Imbens, and Wooldridge (2017), we will not cluster at any higher level of aggregation. We only cluster at the individual level when we pool across waves, as we do in most specifications.

sive margin. The point estimate is 0.004, and is not statistically significant from zero ($p = 0.72, q = 0.45$), compared to a control mean at follow-up of 0.12. This is a tight estimate of a zero effect: we can rule out a 2.5 percentage point increase with 95 percent confidence. Similarly, we estimate a zero effect on hours in self-employment. Our coefficient estimate on profits is also far from significant, but economically fairly large, equivalent to an increase of 12 percent over control group profits.[8]

We find a significant effect of the program on wage-employment — in particular, on the likelihood of having a permanent job, on hours, and on wage income. Each of these outcomes increases by about 8-11% compared to the control group mean. Our estimate on the likelihood of having any kind of wage job is smaller: an increase in 3 percentage points (or 5% in relative terms), statistically significant at 10% even after multiple testing correction. This likely reflects some substitution of casual work for permanent jobs. Franklin (2018) and Abebe, Caria, Fafchamps, Falco, Franklin, and Quinn (2018) report a similar substitution for a transport subsidy intervention in Addis Ababa. We obtain a positive estimate for having a wage job with managerial responsibilities of 1.6 percentage points, relative to a control mean of 11.8% at follow-up. However, this is not statistically significant ($p = 0.15, q = 0.14$).[9]

In addition to these snapshots at six and twelve months, the phone survey data allow us to gain a more detailed understanding of the trajectories of occupation over time. We report these in Figures A.8 and A.9 of the Online Appendix. Consistent with the descriptive data on the rapid labor market entry of the young professionals in our sample, we see a steep and concave trajectory of wage employment for both the treatment and control group. The management experience placement seems to have resulted in a level effect, with no clear effect on the slope: the treatment group has higher employment rates in each month after the placement, though the difference is not always statistically significant for each individual month. Self-employment rates for treatment and control groups are indistinguishable at any point after treatment; this is in line with the regression findings.

Are these results surprising? We measure priors elicited from expert predictions (DellaVigna and Pope, 2018). Before our first public presentation of results from the experiment, we elicited predictions from three distinct groups of experts: international academics working on causal inference, graduate students in development economics at Oxford, and HR experts from Ethiopia. We explained the experiment, showed them the self-employment

[8] The standard error corresponds to 20% of control profits or roughly 9 USD; this is despite winsorizing the top 5% of profits in every wave.

[9] These effects on wage-employment are not driven by interns finding a job at their host firm. At the 12 month follow-up survey, we find only 2 out of 829 interns working at their host firm.

and wage-employment rates at six and twelve months of the control group, and asked for a prediction of the corresponding figures for the treatment group. The results are in Appendix Table A.7. All groups of experts were overconfident in the effect of the program on self-employment. Experts expected a self-employment rate of 16-19% at six months and 17-29% at twelve months, compared to actual rates in the treatment group of 12% and 15%, at the respective points in time. On a whole, academics made the most accurate predictions, with a forecast error of 0-4 percentage points. Academics and students were close with their predictions on wage-employment rates. Ethiopian HR experts overpredicted the effect of the program on self-employment, and underpredicted the wage-employment rates of interns, perhaps even expecting a substitution away from wage-employment. Given their small number ($n = 5$), however, HR expert group predictions have a much higher variance than the other two groups. These result mirror findings from Casey, Glennerster, Miguel, and Voors (2018) that prior beliefs of experts from academia for programs with mixed and modest effects can be reasonably accurate, whereas local experts tend to expect more profound impacts from interventions.

In addition to testing for the effects of the program on the primary labor market outcomes, we evaluate their impacts on several families of secondary outcomes that relate to the primary outcomes, in line with our pre-analysis plan. Specifically, we test for effects of the program on (i) preparations to start a business, (ii) job search, and (iii) business networks. These results are available in Tables A.8 to A.11. First, we test whether being offered a management placement had an effect on preparations to start a business (short of actually starting a business). The evidence in Table A.8 shows that this is not the case.

Second, we test whether treated interns exhibit different job search behaviour, which includes on-the-job search (Appendix Table A.9). We find that they are less likely to be actively searching for a wage job. This is entirely driven by reduced search for a professional wage job. We also find suggestive evidence for higher reservation wages; this is marginally significant but only before multiple-testing correction. Additional analysis suggests that reduced search behaviour is likely driven by a mechanical effect of higher employment rates among the treated, as well as higher satisfaction with existing jobs. When we explore effects separately by follow-up wave (Appendix Table A.10) we find evidence for significantly reduced search only twelve months after treatment, not at six months. This mirrors the pattern over time of employment effects. From the phone survey (Appendix Figure A.10), we find evidence that treated young professionals are more satisfied with their employment situation, although these effects dissipate over time. Third, we test for effects of treatment on business networks. We find some evidence that treated young professionals have a more senior network with business people and managers. However, none of the effects in this family survive multiple-testing correction ($q = 0.15$).[10]

3.2 Management: Attitudes, knowledge and practices

As a second primary outcome, we are interested in whether being placed to shadow a manager had any effect on young professionals' managerial human capital. Measuring this is challenging. Structured management practices can more easily observed at the level of a firm or establishment, as in the seminal work by Bloom and Van Reenen (2007). Yet managerial skills are much more difficult to observe in individuals, especially when those cannot be observed carrying out management tasks. Similar perhaps to driving a car, what matters is the practical application of skills in particular real-world circumstances — not merely theoretical knowledge about rules of the road, or about the correct response in a critical situation.

We approach this measurement challenge in various ways. First, we ask young professionals to self-report their *confidence* in their skills to succeed in various management task. Second, we ask respondents about the *relative* importance of various management practices. This would allow us to at least measure any changes in attitudes. Third, for those young professionals who run their own business at follow-up, we can measure the management practices they apply in their firms; using a survey instrument specifically calibrated to small firms in developing countries by McKenzie and Woodruff (2017).

Young professionals are remarkably confident in their management skills even before they were assigned to our program. On average, respondents at baseline state that they are either confident or very confident in their skills across 10 out of 14 areas of management. Appendix Table A.12 presents the wording of each question. In columns (1) and (2) of Table 3, we report the effect of treatment on two summary measures of management confidence, the sum of categories and a normalized index. We find that treatment significantly increases confidence by either measure.[11] Given their apparent overconfidence, one might ask whether these self-reported outcomes actually pick be any underlying real effect.[12] Reassuringly, we

[10] The effect on networks seems not to be driven by contacts directly acquired through the placement. Only 1 intern lists a contact from their host firm at the 12-month follow-up survey. By contrast, 60 interns list contacts at their current firm. This suggests that the effect on business networks might be at least partly driven by the effect on wage employment.

[11] This test survives correcting for testing of 16 hypotheses in the family, including the individual areas reported in Appendix Table A.13

[12] Indeed, exposure to actual management could have made interns realize that they they were overconfident to start with, and align their expectations accordingly. This is, however, not what we find.

only see average effects on confidence in areas that young professionals were most exposed to in host firms: planning tasks such as cost and demand estimation, or dealing with clients (Appendix Table A.13). We generally see no effects on areas that were off limits to most interns such as dealing with suppliers, and access to finance. This gives us reassurance that, despite the clear overconfidence bias in these measures, respondents do pick up some signal about the effects of the program on perceived managerial skills.

Next, we report differences in rankings across management practices. We anticipated that we would not be able to obtain an accurate measure of knowledge of best practice in management through a survey measure, given that informed respondents would have a good sense that picking the most sophisticated structured practice would be the 'right' answer. We therefore opted to elicit choices about trade-offs by asking young professionals to rank ten 'good' management practices, some of which are more relevant for small firms and some for large firms. Figure 2 shows the distribution of first-ranked practices for the treatment and control group (the sum across categories has to be equal to one). The most important practice for control respondents is separating household and business assets (18%), the least important one is frequently monitoring employee performance (3%). We find that the distributions across groups are significantly different ($p = 0.06$). Interns are more likely to put weight on practices associated with the large host firms they were placed in; in particular sales targets and employee monitoring.

Finally, we report the average effect of treatment on management practices in firms run by respondents. In columns (3)–(7) of Table 3, we report standardized effects on overall practices, and sub-components marketing practices, record-keeping practices, and financial practices, following (McKenzie and Woodruff, 2017). We find some suggestive evidence that placement increased overall management practices. The point estimate is 0.08 standard deviations, and is significant at the 10% level, though only before correcting for FDR. Estimates on individuals components of practices are highly non-significant. By definition, we only observe management practices for those respondents who run a business, so our estimates will reflect a mix of selection and learning about management. To explore this issue further, we split the sample into incumbents – who already had a business at baseline – and entrants (Appendix Table A.14).[13] We find that all the effects are driven by incumbents – the point estimate of treatment on overall practices is 0.16 standard deviations; though again this is only weakly significant. The point estimate for entrants is zero.

[13] This exploratory analysis was not specified in our pre-analysis plan.

3.3 Testing for effects on host firms

We also test for effects on host firms, exploiting the fact that firms were also randomized into participation. We have firm data from a baseline survey, and from a follow-up survey shortly after the conclusion of the program. We again estimate an ANCOVA specification. Since firm randomisation was stratified by gathered fields of available firms, we cluster standard errors at this level.

Hosting a young professional for a short period of time is unlikely to transform how medium and large firms do their business. Nevertheless, the profile of interns is potentially different from that of firms' usual employees, and we anticipated that hosting such interns might change some aspects of firm behaviour. We therefore test for effects on a few selected and pre-specific HR and management outcomes, a few weeks after the placement. We find no effects on firms' advertising, hiring or general management practices (Appendix Tables A.15, A.16, and A.17). We obtain a significantly positive point estimate on separations, which implies that treated firms had an additional 2.9 workers separations (A.16). We find no effect on hiring. However, we cannot rule out that this reflects the mechanical effect from our program, which placed on average 2 interns into host firms. If we assume that all firms include their program interns into the calculation of separations, the coefficient reduces to 1.1 ($p = 0.29$).[14]

4 Heterogeneity in treatment type

By design, the placement experience of interns is heterogeneous, because host firms are drawn from different industries, locations, and sizes. Consequently, the effect of placement on interns may vary with the type of firm to which they are assigned. In particular, the effect of an internship placement may differ with the quality and sophistication of the host firm's management practices. Heterogeneity in treatment type is useful for understanding the mechanisms through which our experiment works. Indeed the average treatment effect could mask substantial difference in how treatment of different types work on the average participant. Some varieties of treatment might be beneficial, while others may be ineffective or harmful. Moreover, there might be match-specific effects of assigning a particular type of treatment to a specific type of subject, in which case the estimated average treatment effect is a function of the particular mechanism by which individuals are assigned to treatment

[14] We conduct a number of additional tests that were pre-specified in our pre-analysis plan, but which we do not discuss in the main text. These can be found in the Online Appendix.

varieties. In such situations, it is therefore important to identify the effect of treatment types on different individuals so as to be able to predict the average treatment of other assignment mechanisms.

In our setting, the allocation of young professionals into batches is random at the margin — after each round of applications, individuals are invited to an induction session in random order. The controlled assignment mechanism lets us determine what assignment would look like under a counterfactual grouping of firms and interms. Simulating other possible groupings allows us to recover the propensity score of each intern-firm match, that is, the conditional probability that a particular intern is assigned to a specific type of host firm. Conditional on this propensity score, the actual assignment of an intern to a particular type of firm is quasi-random: for two individuals with the same propensity score, their actual firm assignment only depends on the particular realization of the grouping of firms and interns into batches. We show in this section how this variation can be used to identify marginal treatment effects — i.e., the causal effect of being assigned to one type of firm over another at each point in the propensity score support. We then use the results to further interpret our experimental findings, and to explore the implications of alternative, counterfactual assignment rules that might be of interest to policymakers. We also discuss the applicability of the method to other assignment mechanisms used in practice.

4.1 Identifying causal effects under controlled assignment

Denote the rankings of firm f over a set of interns I with the vector $\boldsymbol{r}_{fI}$. We treat $\boldsymbol{r}_{fI}$ as a draw from a probability mass function given by the vector-valued functional $\boldsymbol{\rho}_f$, which takes as its inputs the observable intern characteristics for the set of interns considered by firm f (denoted $\boldsymbol{w}_1$,$\boldsymbol{w}_2$,...):

$$\boldsymbol{r}_{fI} \sim \boldsymbol{\rho}_f(\boldsymbol{w}_1,\boldsymbol{w}_2,...). \tag{2}$$

Note that we index the functional by f; this allows that different firms have different preferences over various intern characteristics. For notional brevity, we use the upper-case F to denote a given set of firms, and use $\boldsymbol{\rho}_F$ to denote the set of functionals relating to that given set of firms. Note that, where the distribution implied by $\boldsymbol{\rho}$ is non-degenerate, the same firm has idiosyncratic variation in expressed rankings, even after conditioning on intern characteristics. In due course, we use a flexible semi-parametric model that generates both across-firm variation and also idiosyncratic variation of each firm's rankings; this will imply a specific and tractable form for $\boldsymbol{\rho}_f$, but none of the reasoning in this section depends upon that model.

Symmetrically, the realisation of interns' rankings over firms (generically, $\boldsymbol{r}_{iF}$) is drawn from a probability mass function given by the functional $\boldsymbol{\tau}_i$, where observable firm characteristics are denoted by $\boldsymbol{x}$:

$$\boldsymbol{r}_{iF} \sim \boldsymbol{\tau}_i(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots). \tag{3}$$

The matching function ψ takes as inputs both sets of rankings — firms over interns, and interns over firms — in a given batch. It has as an output the assignment matrix $\boldsymbol{m}_{IF}$:[15]

$$\boldsymbol{m}_{IF} = \psi([\boldsymbol{R}_{IF}, \boldsymbol{R}_{FI}]). \tag{4}$$

This implies that the assignment mechanism is deterministic: each intern is offered a particular place with certainty. Our deferred acceptance mechanism with no ties in rankings is an example of a deterministic mechanism.[16] It is further helpful to denote the assignment of a particular intern i to a particular firm f as element (i,f) of equation 4:

$$\boldsymbol{m}_{if} = \psi_{if}([\boldsymbol{r}_{iF}, \boldsymbol{R}_{-i,F}], \boldsymbol{R}_{FI}). \tag{5}$$

Here our notation separates $\boldsymbol{R}_{IF}$ into the ranking intern i gave to firms, and the rankings from all other interns in the batch, whom we denote by $-i$.

The key to using our assignment mechanism for causal inference is to recognize that one of the variables that drive this deterministic assignment can be viewed as the realisation of a random variable. If we then integrate out this random variable (that is, remove the conditioning upon its realisation), we have a stochastic mechanism that gives rise to a well-defined propensity score. The random variable in our field experiment is the group composition. In particular, young professionals were invited to batches randomly; therefore the other group members, $-i$, are exogeneous in their characteristics and preferences to

[15] We adopt the convention that rows represent interns and columns represent firms. The dyadic ranking matrix $\boldsymbol{R}_{IF}$ then stacks the row vectors $\boldsymbol{r}_{iF}$ into an $|I| \times |F|$ matrix; $\boldsymbol{R}_{FI}$ stacks the column vectors $\boldsymbol{r}_{fI}$. The assignment matrix $\boldsymbol{m}$ is a $|I| \times |F|$ matrix, where $|I|$ is the number of interns of the batch and $|F|$ the number of firms. Each row of $\boldsymbol{m}$ contains exactly one element which is equal to one, else zero; and for each column $\sum_i \boldsymbol{m}_{If} = \kappa_f$, firms f's capacity. We do not restrict capacity to be equal to one; hence a firm can offer multiple places.

[16] More generally, of course, *any* mechanism can be considered as deterministic if we condition on any random variables that are used for tie-breaking (such as lottery numbers in school choice mechanisms).

i, the intern whose outcomes are of interest to us.[17]

From these primitives, we can write the probability of a given intern (i) being matched to a given firm (f), conditional on the rankings and characteristics of that intern, and conditional on the characteristics and preferences over the set of firms in the batch. This simply involves integrating equation 5 over the joint distribution of characteristics and preferences of other potential interns who could have joined i's batch:[18]

$$\begin{aligned} q_{if} &\equiv \Pr(m_{if} = 1 | \boldsymbol{r}_{iF}, \boldsymbol{w}_i, \boldsymbol{\rho}_F, \boldsymbol{X}_F) \\ &= \int \psi_{if}\Big(\big[\boldsymbol{r}_{iF}, \boldsymbol{R}_{-i,F}(\boldsymbol{X}_F)\big], \boldsymbol{R}_{FI}\big([\boldsymbol{w}_i, \boldsymbol{W}_{-i}]\big) \Big) \, d\mathcal{F}_{(\boldsymbol{W}_{-i}, \boldsymbol{\tau}_{-i})}. \end{aligned} \tag{6}$$

Note that there is an important asymmetry in equation 6 with respect to the information that we condition on for different agents. For intern i — the intern whose assignment probability we focus on — we directly condition on their *observed rankings* of host firms. We also condition on i's characteristics that the firms observe. For firms, we condition on their characteristics and *preferences* — that is, we condition on the primitives of their choices, rather than observed choices themselves.

In our analysis, we focus on i's assignment probability to a particular *type* (or variety) of firm. To fix ideas, denote $\tilde{D}_f$ as a dummy for whether firm f has an above-median score on our measure of structured management practices; denote D_i as a dummy for whether intern i is assigned to such a firm. Then we obtain the following conditional probability:

$$\Pr(D_i = 1 | \boldsymbol{r}_{iF}, \boldsymbol{w}_i, \boldsymbol{\rho}_F, \boldsymbol{X}_F) = \sum_{f=1}^{F} q_{if} \cdot \tilde{D}_f \equiv p_i. \tag{7}$$

We call expression 7 the *i-conditional propensity score*, because it fully conditions on every aspect of i that matters for i's assignment probability. By construction — and by virtue of our programme design, which strictly assigns placement with the DA mechanism — rankings are the only way that aspects about i that are not in $\boldsymbol{w}_i$ can matter for assignment. In other words, observed rankings carry all the information about how unobservables (such as preferences) matter for selection of i into a particular type of firm assignment. Further,

[17] We note that in principle, firm characteristics and preferences could play a similar role as a random variable. However, as we noted earlier in our discussion of implementation, firms were scheduled purposefully, so we prefer to condition on instead of integrating out their characteristics and preferences. This mode of implementation likely extends to other settings where individuals on the one side are matched to institutions on the other side; where institutions are firms, training providers, or consultants, to give a few examples.

[18] For compact notation, $\boldsymbol{X}_F$ stacks all the characteristics of the $|F|$ firms, and $\boldsymbol{W}_{-i}$ and $\boldsymbol{\tau}_{-i}$ respectively stack characteristics and preferences of interns other than intern i.

the characteristics $\boldsymbol{w}_i$ are the only aspect about i that enter the firms' rankings and hence their choices. This implies that there is nothing unobservable about i that will correlate with assignment to a high-management firm.

This can be formalised through an important property of controlled assignment mechanisms: the 'equal treatment of equals'. In our context, this property can be stated as following:

Definition 1 ***Equal treatment of equals*** *('ETE')*
Consider two individuals i and j, such that $r_{iF} = r_{jF}$ and $\boldsymbol{w}_i = \boldsymbol{w}_j$, and two identical batches of firms, which implies that $\boldsymbol{X}_F$ and $\boldsymbol{\rho}_F$ are fixed. We say that mechanism ψ satisfies the equal treatment of equals property if these individuals have the same match probabilities for all firms in the batch: $q_{if} = q_{jf}$ $\forall f$. (Note that, in our context, ETE implies trivially that individuals have the same i-conditional propensity scores: $p_i = p_j$.)

In our mechanism, it is sufficient for ETE to hold that:

$$Y_{i0}, Y_{i1} \perp\!\!\!\perp (\boldsymbol{w}_k, \boldsymbol{\tau}_k) \,|\, \boldsymbol{Z}_{iF} \qquad \forall\, k \neq i, \tag{8}$$

where Y_{i0} and Y_{i1} are potential outcomes for young professionals (such as their labor market status) under $D_i = 0$ and $D_i = 1$ respectively. By our definition of ETE, the conditioning vector $\mathbf{Z}_{iF} = \{r_{iF}, \boldsymbol{w}_i, \boldsymbol{X}_F, \boldsymbol{\rho}_F\}$. Equation 8 therefore states that characteristics and preferences of other interns are independent of potential outcomes for intern i, after conditioning on intern i's rankings and observable characteristics, and after conditioning on the characteristics and preferences of firms in the batch.

One way to meet this condition is through random assignment of interns to batches. Abdulkadiroğlu, Angrist, Narita, and Pathak (2017) show a assignment by a stochastic mechanism which obeys ETE implies ignorability of treatment assignment. In our case, 'treatment' is defined as the firm type that an intern is assigned to. That is, conditional on $\mathbf{Z}_{iF}$, assignment to a high-type firm is independent of potential outcomes:

$$\Pr(D_i = 1 | \mathbf{Z}_{iF}, Y_{i1}, Y_{i0}) = \Pr(D_i = 1 | \mathbf{Z}_{iF}). \tag{9}$$

In other words, conditioning on all variables $\mathbf{Z}_{iF}$ that pertain to intern i's probability of selection into being placed in a high-type firm eliminates selection bias. Conditional on $\mathbf{Z}_{iF}$, selection in placement with a high-type firm is quasi-random — and, in particular, is independent of potential outcomes for i. (To illustrate, ETE might be violated in our context if, for example, groups of friends had been allowed to sign on to attend the same batch together, or if interns had collaborated to formulate their rankings; neither of these possibilities were allowed by our design.)

The final step for identification of the causal effect of D on interns' outcomes is a straightforward application of the propensity score theorem: Rosenbaum and Rubin (1983) show that if ignorability holds,[19] then conditioning on the propensity score is sufficient for eliminating selection bias. That is, instead of conditioning on $\mathbf{Z}_{iF}$ it suffices to condition on $p_i = \Pr(D_i = 1 | \mathbf{Z}_{iF})$ to ensure that D_i is conditionally independent of potential outcomes:

$$\Pr(D_i = 1 | p_i, Y_{i1}, Y_{i0}) = \Pr(D_i = 1 | p_i). \tag{10}$$

That is, since we control the DA matching algorithm *and* the information set that firms and interns use to produce their ranking, $\mathbf{Z}_{iF}$ is the complete set of variables that determines assignment, as in equation 7. In other words, by design of our assignment mechanism, once we condition on the propensity score, we do not have any other unobservables which could correlate with assignment to a particular type of firm.

We note that this result does not rely on the particular mechanism used (in this case, a deferred acceptance algorithm). Any mechanism that is either stochastic by design, or which can be interpreted as a particular realization of a stochastic process, will fit into this framework. For instance, when a matching treatment is carried out with a random serial dictatorship (as in Breza and Chandrasekhar (2019)), our framework could in principle be applied. An important consideration for practical applications is that the mechanism needs to be reproducible for a different draw from the underlying stochastic distribution. We now turn to resolving this issue for our context.

4.2 A Bayesian estimator of the propensity score

The previous section showed how, in a controlled assignment mechanism such as DA, the propensity score can be calculated by simulation: numerically integrating out the random variable involved in assignment. By simulating many counterfactual group compositions, we obtain the propensity that an intern was assigned to their actual host firm when the experiment was actually run. In our application, this requires integrating over the joint distribution of other interns' characteristics and preferences ($\mathcal{F}_{(\boldsymbol{W}_{-i}, \boldsymbol{\tau}_{-i})}$), and the preferences

[19] An additional common support condition is also required; this is met straightforwardly in our case.

of firm over simulated interns ($\boldsymbol{\rho}_f$).

In principle, several different methods could be used to construct these objects; none of our earlier reasoning depends upon the particular estimation method used to characterise participants' preferences. In our application, we use a Bayesian classification algorithm to train a generative statistical model of heterogeneous preferences for interns and firms. Bayesian classification models are a standard element in the toolbox of machine learning algorithms.[20] An advantage over more agnostic classes of algorithms — such as random forests or neural networks — is that we can specify a statistical model that fully takes into account the rank-ordered nature of our data, and which has a straightforward economic interpretation in terms of characteristics and preferences. Most importantly, such a model is generative: it allows us draw from the underlying posterior probability distribution, which then allows straightforwardly for numerical integration of the kind required by equation 6. We discuss our model — namely, a Plackett-Luce model nested in a discrete finite mixture framework — in Appendix B.

With Bayesian estimates in hand, it is straightforward to draw from the posterior distribution of p_i (as defined in equation 7); we do this by replacing the various unknown distributions in equation 7 by their estimated posterior distributions. First, we draw from the marginal distribution of $\boldsymbol{W}_{-i}$ by using a discrete non-parametric density estimate over intern characteristics; specifically, we treat $\boldsymbol{w}_j$ as a high-dimensional categorical variable, and apply a weak Dirichlet prior across all possible categorical realisations. To implement, we use the observed characteristics of interns in the given batch, and in batches occurring at similar times.[21] We further rely on assumption 8 which implies independence across individuals; this simplifies the joint distribution $\mathcal{F}_{(\boldsymbol{W}_{-i},\boldsymbol{\tau}_{-i})}$ to a product of densities $\mathcal{F}_{(\boldsymbol{w}_j,\tau_j)}$ for each individual $j \in -i$. For each given draw of $\boldsymbol{w}_j$, we draw from the posterior of the Plackett-Luce parameters and the discrete finite mixture distribution. Finally, we draw independent idiosyncratic taste shocks. For each $j \in -i$, we form simulated utilities, order those, and record the resulting ranking of firms: together, these steps constitute a single draw from the posterior distribution $\hat{\mathcal{F}}_{(\boldsymbol{W}_{-i},\boldsymbol{\tau}_{-i})}$.

Equation 7 conditions on the set of preferences for all firms in the batch: $\boldsymbol{\rho}_F$. Since the firms are held fixed, to draw from the posterior distribution of this object, we can condition on firm characteristics $\boldsymbol{x}_f$, and need only to estimate the conditional distribution of each firm's preferences, given the firm's recorded rankings and characteristics: $\hat{\mathcal{G}}_{\boldsymbol{\rho}_f|\boldsymbol{R}_{fI},\boldsymbol{x}_f}$. This

[20] See, for example, the standard graduate level introductory textbook into machine learning by Bishop (2006).

[21] We use a bandwidth of 2 batches; thus, for example, for batch 20, we use intern data from batches 18 to 22. We use a Dirichlet prior of 0.0025.

obtains directly from the posterior distribution of our previously estimated Plackett-Luce model of firm preferences. Specifically, we draw from the posterior distribution of the Plackett-Luce parameters. For each draw, we calculate the likelihood of a given firm's observed ranking and given the firm's characteristics; by Bayes Rule, this implies a conditional probability that the firm belongs to each estimated type. We simulate by drawing from this conditional probability; as in the case of interns, we then form simulated utility, which we rank. This generates a single draw from the posterior $\hat{\mathcal{G}}_{\boldsymbol{\rho}_f|\boldsymbol{R}_{fI},\boldsymbol{x}_f}$.

For each draw from $\hat{\mathcal{F}}_{(\boldsymbol{W}_{-i},\boldsymbol{\tau}_{-i})}$ and $\hat{\mathcal{G}}_{\boldsymbol{\rho}_f|\boldsymbol{R}_{fI},\boldsymbol{x}_f}$, we form simulated assignments $\boldsymbol{m}_{if}$. By drawing repeatedly, we apply the integral described in equation 6; the propensity score, p_i follows straightforwardly by equation 7. The posterior distribution of p_i is formed simply by looping repeatedly over this process. (For all of the analysis that follows, we conduct inference by allowing both for parameter uncertainty conditional upon p_i, and for uncertainty in p_i itself; we do this using a method broadly analogous to the 'combining rules' of Rubin (2004).)

4.3 Validating the estimated propensity scores

There are two ways of checking the constructed propensity scores: a posterior predictive check, and a baseline balance check. In this section, we discuss both.

First, we conduct a posterior predictive check. To do this, we compare the probability of a given firm-intern match with the constructed probability q_{if}. This checks the identity in equation 6, namely that $q_{if} \equiv Pr(m_{if} = 1 \mid \boldsymbol{Z}_{iF})$. Applying the propensity score theorem, this becomes $q_{if} = Pr(m_{if} = 1 \mid q_{if})$. In other words, we are checking, at the dyadic intern-firm level, whether the mean assignment probability in the data corresponds to the simulated assignment probability. For example, of those firm-intern dyads having $q_{if} = 0.8$, we expect 80% actually to have been matched.[22] Specifically, we run a non-parametric regression of a dummy for a match between intern i and firm f, explained by our constructed probability q_{if}; in principle, this should approximate a 45-degree line through the origin.[23]

We plot the results in Figure 3. We note two findings. First, we find that the estimated curve indeed lies very close to the 45-degree line across the entire support of the simulated match probability, which comprises almost all of the unit interval. We only have few obser-

[22] A similar check applies for the propensity score: for interns with $p_i = 0.8$, we expect on average 80% to be matched to a high-management firm. The dyadic-level check that we perform implies (but is stronger than) this check on the propensity score, given the propensity score aggregation in equation 7.

[23] Specifically, we do this using a kernel regression, with a log transformation of the simulated probability of match; we use a bandwidth of 0.1 in that log space.

vations with simulated match probabilities above 0.6; the estimated relationship is therefore noisier at the upper end of the distribution but still follows the 45-degree line closely. Second, we find that our simulated match probability covers the full range of the unit interval.[24]

As a second check, we examine whether controlling for the simulated propensity score — the match probabilities aggregated according to equation 7 — achieves baseline balance.[25] We do this with a series of estimations, all reported in Table 4. In Panel A of Table 4, we show a simple OLS regression of baseline outcomes on a dummy indicating a high-management firm, without any controls. We see that interns assigned to a high-management firm are positively selected. They are more likely to be in a (permanent) job at baseline, work more hours and earn a higher wage income. All of these outcomes are significant at 10%. More suggestively, interns matched with high-management firms are also more likely to be self-employed; although the coefficients are smaller and not statistically significant. These results are important for motivating our propensity score analysis; they show empirically why assignment to high-management firms is clearly not 'as if random'.

We then condition on the simulated propensity score, in several ways. We first introduce a control function of p into the OLS regression, of multiple functional forms: linear in panel B; a saturated model of centile dummies in panel C; and a semi-parametric regression in panel D.[26] Second, we implement propensity score conditioning by applying inverse probability weighting to the observations in either category of the dummy. We find that, after conditioning on the propensity score, for each regression model that we try, none of the baseline occupation and earnings variables are significantly different for interns assigned to a high-management firm. The coefficients are always closer to zero than in Panel A. This evidence shows that propensity score conditioning is effective in controlling for the selection into high-management firms.

[24] If approximation error gave rise to a serious distortion, then we might have found that our simulated match probability covered a shorter range; this is not the case here.

[25] Our sample here — and for subsequent analysis — includes all treated individuals with propensity scores strictly between zero and one, for whom the common support condition holds.

[26] We also explore higher-order polynomials in Appendix Table A.18. Results are almost identical to controlling linearly for the propensity score. This is unsurprising, given that the Härdle-Mammen tests in panel D pass comfortably.

4.4 Marginal treatment effects

To understand the heterogeneous effect of our different varieties of treatment, we calculate the marginal treatment effect ('MTE'):

$$MTE(p) = \frac{\partial E(Y|P(\boldsymbol{Z}_{iF}) = p)}{\partial p}, \tag{11}$$

where p is a realisation of the propensity score $P(\boldsymbol{Z}_{iF})$ (Heckman, 2005; Heckman and Vytlacil, 2007; Carneiro, Heckman, and Vytlacil, 2011). The conditional expectation of Y given $P(\boldsymbol{Z}_{iF})$ is then mechanically given by integrating up:

$$E(Y|P(\boldsymbol{Z}_{iF}) = p) = \mu_0 + \int_0^p MTE(u_s) du_s, \tag{12}$$

where μ_0 is a constant of integration that can be interpreted as the average potential outcome in the group exposed to the numeraire type. The 'treatment effect' in our setting is the differential effect of being assigned to the treatment type high-management host as opposed to a low-management host firm.

An alternative and equivalent formulation of the conditional expectation of Y given $P(\mathbf{Z})$ is given by:

$$E(Y|P(\boldsymbol{Z}_{iF}) = p) = y_0(p) + p \cdot (y_1(p) - y_0(p)). \tag{13}$$

To estimate the MTE, we run a kernel regression of outcomes on our empirical propensity score, the simulated assignment probability. We do this separately for those interns assigned to a high-management firm, and those assigned to a low-management firm. The difference between these two curves $(y_1(p) - y_0(p))$ is the effect of being assigned to a high-management host firm rather than to a low-management host, conditional on a given value of p (that is, the integral of $MTE(p)$ over a small interval in the p-space).[27]

Our primary outcome of interest for this analysis is total income — that is, the sum of income from self-employment and income from wage employment (where, to avoid selection effects, this is set to zero for individuals who are not employed). In Panel A of Figure 4, we show the estimated outcome for a high-management firm (solid, in red) and for a low-management firm (dotted, in blue); the shaded regions show 90% confidence intervals. (As in our main analysis, these effects are estimated by pooling data from the

[27] We present all our analysis of heterogeneity in treatment type in the form of graphs that show the MTE. In Appendix C, we additionally provide regression evidence on the average differential effect of being placed into a high-management firm as opposed to a low-management firms.

six-month and 12-month follow-up surveys.)

The results are striking. First, consider interns assigned to low-management host firms. For such participants, follow-up income is remarkably stable, regardless of the propensity score: expected monthly follow-up income is approximately 3500 birr, across the range of p_i (that is, about 125 USD). For interns assigned to high-management host firms, however, the story is very different. Interns with a low propensity score do very poorly if assigned to high-management firms: for interns with p_i close to 0, for example, the expected follow-up monthly income is just 2200 birr (about 75 USD). In income terms, at least, these interns would be much better placed with low-management hosts. However, in contrast to the low-management graph, this plot has a steep slope in p_i — such that, in expectation, an intern having $p_i \approx 1$ will, if placed with a high-management host, earn about 4600 birr at follow-up (that is, about 165 USD).

We can disaggregate this effect between self-employment profit and wage income. Panel B of Figure 4 shows the equivalent graph as Panel A, drawn for profit income alone. The graph shows a very clear pattern, matching the general shape of Panel A. In contrast, the equivalent figure for wage income (Figure A.12, in the Online Appendix) shows no such pattern: there, we find a general increase in income with p_i, but not striking differences between slopes for those assigned to low-management hosts and those assigned to high-management.

In Panel C, we delve further — by testing the effect on probability of self-employment. Here, too, we see interesting heterogeneity — both by p_i and by host management quality. For interns with a low propensity score, the probability of self-employment at follow-up is very similar between those assigned to low-management hosts and those assigned to high-management hosts. As p_i increases, however, this story changes dramatically: for interns with p_i close to 1, assignment to a low-management host leads to a self-employment probability of about 8%, whereas assignment to a high-management host leads to a self-employment probability more than twice as large (about 17%). This change in relative magnitudes is broadly similar to the change in relative magnitudes observed in Panel B; thus, it seems that the striking pattern in Panel A is driven by a differential effect of high-management firms encouraging interns with high propensity scores to move into (profitable) self-employment.

More generally, these findings are a key ingredient to fully understanding the program's impact on occupation and employment. Recall that, overall, we have found that the program had a zero *average* effect on self-employment. Our analysis of MTE now suggests that this average effect masks substantial heterogeneity. In particular, we find large and positive effects on self-employment for individuals who were most likely to be assigned to a firm from the upper half of the management distribution, and who were actually assigned to such a firm. This is offset by low-p individuals, and high-p individuals not assigned to a high-management firm; both of these groups are less likely to be self-employed. The latter group experiences the biggest increase in wage-employment, which is indicative of occupational substitutions depending on the kind of host firm. Nevertheless, wage-employment increases for high p-individuals for any kind of host firm. Indeed, the MTE results suggest that the overall effect of the program in increasing wage-employment is entirely driven by individuals with higher p.

4.5 Alternative implementations of propensity score simulation

As discussed earlier, none of our reasoning on identification in section 4.1 depends on the precise nature of the assignment mechanism used, nor does it demand any particular method to recover the propensity score in an empirical application. We now briefly discuss two alternative approaches for the latter, before moving on to consider the effects of alternative assignment mechanisms.

4.5.1 Drawing from the empirical distribution of observables

In our preferred implementation earlier, we draw counterfactual intern rankings from (i) an estimated model of preferences, and (ii) an estimated distribution of observable characteristics that is estimated locally for each batch. This local density estimation approach is our preferred method because it allows for the distribution of characteristics to change across time.[28]

We now consider a simpler alternative: draw the characteristics of other interns in the batch from the empirical distribution of interns across *all* batches of our experiment, with replacement. We produce the equivalent to the baseline balance regression from Table 4 and the estimated match probability at each level of the propensity score from Figure 3 in the Online Appendix (Panels A of Table A.19 and Figure A.13). We find that both of these results are essentially unchanged: the relation between the estimated probability of a match and the simulated propensity score still resembles a 45 degree line; and conditioning on the propensity score achieves baseline balance on outcomes in a similar way to before. (Indeed, there is a correlation of 0.99 between at the values of q_{if} obtained using our original method and those obtained using this alternative method.) From this we conclude that, while we

[28] See Abebe, Caria, and Ortiz-Opsina (2018) for evidence on how the average quality of job seekers of a given cohort deteriorates over time, for a comparable sample of young job seekers in Addis Ababa; this result might generate a concern that individuals joining the experiment in the latter batches are not comparable to those joining at the outset.

prefer on conceptual grounds a method that is robust to a shifting composition of intern characteristics, in practice this makes no meaningful difference in our application.[29]

4.5.2 Drawing from the empirical distribution of rankings

Another alternative that naturally offers itself is to simulate the counterfactual distribution by repeatedly drawing, with replacement, from the observed distribution of rankings — so that, when an intern is resampled, we take both her rankings and her relative location in the rankings of the host firms. If successful, this alternative would remove the need for a generative preference model. This approach is intuitively appealing, but also has some undesirable features. Specifically, this method samples from the vast space of possible rankings simply by replicating a small number of observed rankings; in that sense, it suffers from an 'unobserved species' problem, by refusing to countenance the possibility of preference orderings different from those already observed in the data (Efron and Thisted, 1976). Further, and closely related, this implies that every simulation draw using this method looks very different to the data in one key respect: the within-batch correlation of rankings. Across our 42 batches, we never observed any case in which two firms gave identical rankings of interns, nor any case in which two interns gave identical rankings of firms; however, such ranking duplications occur on every single draw of this proposed method.

Empirical simulation results confirm these concerns. We find that the relation between the simulated assignment probability and the estimated probability of a match (in Panel B of Online Appendix Figure A.13) follows the shape of a sigmoid, instead of a 45-degree line. This implies a compression of simulated assignment probabilities, such that they are understated for high-p individuals and overstated for low-p individuals. Based on this empirical failure, we therefore cannot rely on ranking-resampling to generate propensity scores — at least in this context.[30]

[29] We also experimented with drawing from the estimated distribution for the entire sample — *i.e.* by increasing the bandwidth of our density estimation to all batches — and from the empirical distribution of only the batch itself. Results are almost indistinguishable from our main results.

[30] It is worth noting that the propensity score generated by this method appears to be a monotonic transformation of the intended propensity score; therefore, while this score could not be used to calculate marginal treatment effects, conditioning on the propensity scores based on the bootstrap does achieve baseline balance in a similar way that other approaches do; see Table A.19, Panel B.

5 Counterfactual mechanism design: A bivariate propensity score method

In the previous section, we calculated marginal treatment effects of assignment to a high-management host (as opposed to a low-management host), in order to understand the mechanisms driving our experimental results. We now extend this analysis in order to predict likely effects under alternative mechanisms. Our basic insight here is quite simple: *the generative model that allows us to construct a propensity score for our implemented mechanism can also generate propensity scores for any other controlled mechanism that relies only on rankings as its input*. Formally, we do this by holding fixed the estimates from our preference model, and modifying the matching function ψ in equation 4. We now denote the propensity scores under our implemented mechanism as p_a; we denote the propensity scores under some alternative controlled mechanism as p_b.[31]

With p_a and p_b in hand, we can estimating the expected outcome under an alternative mechanism b simply by a weighted integration:

$$E_b(Y)=\int\int[p_b\cdot y_1(p_a,p_b)+(1-p_b)\cdot y_0(p_a,p_b)]\,f(p_a,p_b)\,dp_a dp_b, \tag{14}$$

where $y_1(p_a,p_b)$ and $y_0(p_a,p_b)$ respectively denote average outcomes for individuals assigned to high-management and low-management firms (estimated here using bivariate kernel regression). This follows from the representation of the conditional expectation of Y under the implemented mechanism in equation 13. The weighting here by p_b and $1-p_b$ reflects the use of the alternative mechanism; if we were to weight instead by p_a and $1-p_a$, we would recover the estimated marginal treatment effects discussed in the previous section.[32] Note that random matching can be evaluated as a degenerate special case of this

[31] Note that this approach assumes that, both under our implemented mechanism and under any candidate alternative mechanism, both firms and individuals report truthfully about their preferences — notwithstanding that, for a deferred acceptance algorithm, truthful reporting is incentivized only for the proposing side (see, for example, Roth and Sotomayor (1990)). We view this assumption as reasonable in this context given the substantial uncertainty that both firms and individuals face about their batch. Specifically, the firms did not know the identities of the other firms in their batch, and do not know the applicant individuals except for their stylized CVs; conversely, the individuals did not know each other when making their rankings, and were not allowed to communicate to coordinate their rankings. For this reason, regret-aversion is likely to encourage participants to report truthfully: Fernandez (2018). Put differently, it is not at all clear how one of our individual applicants *should* have deviated from truth-telling given the heterogeneity and complexity of the decision environment — even if such deviations are theoretically feasible. Finally, we note that the estimates obtained in Appendix B (from our Plackett-Luce model) are entirely consistent with the kind of preferences that different applicants would be likely truthfully to hold.

[32] Similarly, we could also obtain the same result by evaluating equation 14 for the degenerate case where the alternative mechanism is identical to the mechanism actually used: $p_b\equiv p_a$.

expression, in which p_b is fixed for all individuals in the sample.[33]

Our estimation of expected outcomes under an alternative mechanism builds on the result that any treatment effect of interest can be represented as a suitably weighted integral over the MTE (Heckman, 2005; Heckman and Vytlacil, 2007). In particular, our notion of counterfactual expected outcomes of alternative mechanism is closely related to the policy-relevant treatment effect (PRTE) proposed by Heckman and Vytlacil (2001) and Heckman (2005). Estimation of 14 necessitates the same assumptions that are detailed in Heckman (2005) for the PRTE. In particular, we need to assume that alternative policies only change the assignment probabilities but not the treatment effects themselves. In our setting of alternative assignment mechanisms, this assumption is straightforwardly met.

We apply our method using five feasible mechanisms: (i) random matching, (ii) firm-proposing DA (that is, the mechanism actually implemented in our experiment), (iii) intern-proposing DA, and then two forms of random serial dictatorship (Abdulkadiroğlu and Sönmez, 1998, 1999): (iv) intern-proposing RSD and (v) firm-proposing RSD. As a benchmark, we compare our estimates to estimates from an infeasible social planner — who knows the form of $y_1(p_a,p_b)$ and $y_0(p_a,p_b)$, and who chooses an assignment function $a:(p_a,p_b)\rightarrow\{0,1\}$ in order to maximise the expected outcome, subject to the capacity constraint implied by the number of places at high-management firms.[34]

In Figure 5, we illustrate the three bivariate propensity score distributions: we compare propensities under firm-proposing DA to propensities under intern-proposing DA (Panel A), intern-proposing RSD (Panel B), and firm-proposing RSD (Panel C). The general patterns are clear. Relative to firm-proposing DA, intern-proposing DA is more extreme at the limits: individuals with a propensity score close to 1 see an increase in their propensity score when moving from firm-proposing DA to intern-proposing DA, and individuals with a score close to 0 see a decrease. Under intern-proposing RSD (Panel B), this same pattern is even more extreme. Finally, under firm-proposing RSD, the pattern looks much closer

[33] Thus, for example, if 50% of host firms are 'high-management' and 50% are 'low-management', random matching could be evaluated by setting $p_b = 0.5$ in equation 14.

[34] That is, the infeasible planner solves:

$$\max_{a(p_a,p_b)\in\{0,1\}}\int\left[a(p_a,p_b)\cdot y_1(p_a,p_b)+(1-a(p_a,p_b))\cdot y_0(p_a,p_b)\right]f(p_a,p_b)\,dp_a dp_b,$$
$$\text{subject to}\int a(p_a,p_b)f(p_a,p_b)\,dp_a dp_b=\bar{a},$$

where $\bar{a}$ is the proportion of places that are in high-management firms. When implementing this infeasible planner, we select p_b as the propensity under intern-proposing RSD; this maximises the planner's expected outcomes, relative to choosing p_b as the propensity from any other mechanism considered.

to randomness.

In Figure 6, we show the consequences for estimated total employment income under our various potential mechanisms. The solid bar at the top shows the expected effect had we used random matching rather than a DA algorithm; we estimate that those assigned to the management placement would, on average, have enjoyed only a modest increase in average total monthly income relative to the control group (namely, an increase of about 130 birr). The figure shows that, by using firm-proposing DA, we increased this substantially: by about another 250 birr per month, or about 48% of the difference between random matching and the infeasible planner, and about 15% of average control group income from employment.[35]

More generally, the pattern under other potential mechanisms in Figure 6 is clear: had we used firm-proposing RSD, we would have closed about 30% of the gap to the planner; had we used intern-proposing DA, we would have closed about 62%; and had we used intern-proposing RSD, we would have closed about 73%. In comparison to firm-proposing DA, we find significant differences to random assignment (significant at the 99% confidence level), to the intern-proposing DA algorithm (significant at the 99% level) to the intern-proposing RSD algorithm (significant at the 90% level), and to the infeasible planner (significant at the 99% level).[36]

This result bridges an important conceptual gap: between interns' expressed ranking preferences, and their potential empirical outcomes from matching. It is well known theoretically that, under truth-telling, intern-proposing RSD is Pareto-efficient for interns, intern-proposing DA is the best stable assignment for interns, firm-proposing DA is the worst stable assignment for interns, and firm-proposing RSD is valuable to interns only insofar as firm preferences happen to correlate to intern welfare (Roth and Sotomayor, 1990). However, for a policymaker who cares about a particular empirical outcome — for example, as here, total employment income — it is not clear how ranking preferences map into desired outcomes. It may be, for example, that interns know little about the likely outcomes under alternative potential hosts. In that case, a policymaker might prefer a paternalistic approach, choosing to implement firm-proposing RSD or firm-proposing DA on the basis that 'firms know best'. Indeed, this was part of our own philosophy in choosing to use firm-proposing RSD to conduct the experiment: we had anticipated that established firm managers would be more effective than young interns at predicting effective matches. The results in Figure 6 reject this philosophy emphatically. They show that the pattern of potential earnings

[35] This difference is significant at the 99% confidence level.

[36] These results may seem incongruous given the overlapping of the 95% confidence intervals shown in Figure 6; this is explained by a high positive covariance of estimates between these various objects.

across mechanisms matches exactly the pattern that theory predicts for 'welfare' — showing empirically that intern earnings are improved by mechanisms that empower interns.

These findings have important implications for the design of field experiments. First, we show that the mechanism for assigning participants to treatment varieties *within* the treatment group matters profoundly. Indeed, our results in Figure 5 suggest that — within the finite set of mechanisms that we consider — total income gains relative to the control group vary substantially: between about 120 and 670 Ethiopian birr, compared to a control mean of about 3,400 birr. Second, we demonstrate how assignment of varieties by one controlled assignment mechanism allows the researcher, through the MTE, to estimate counterfactual outcomes under a whole set of assignment mechanism that rely on the same primitives (*i.e.* rankings). This obviates the need for further experimentation on the same population and thus addresses important concerns about the burden that experimentation potentially places on participants (Narita, 2018) (to say nothing of the massive practical impediments to testing each alternative mechanism through a large-scale field experiment).

In this context, we have applied our method to understanding the effect of alternative matching algorithms. However, our approach could readily be adapted to measure effects of a wide variety of other aspects of mechanism design — for example, one could use this approach to measure the effects of market thickness (by, for example, simulating propensity scores under smaller batch sizes), or to measure the effects of pre-sorting applicants (by, for example, having some batches dedicated to manufacturing firms and interns with technical training), or to measure the effects of inviting surplus firms, some of which then remain unmatched.[37]

6 Conclusion

In this paper we have reported the results from a novel experiment on management experience. We work with a population that is highly relevant for studying this topic: educated labor force entrants with a revealed aspiration to become a manager or an entrepreneur in a rapidly growing African city. We randomly assign half of our sample to a one-month placement in an established firm where the intern shadows a middle manager. Debriefing interviews with firms and interns suggest that the placement largely worked as intended. We find that the intervention does not increase realized or planned self-employment, al-

[37] In the matching literature, such unmatched firms are sometimes referred to as 'lone wolves'; see McVitie and Wilson (1970); Roth (1984).

though it does increase confidence in own managerial ability. Treated individuals are instead more likely to aspire to a better wage job and to find a more stable job with a permanent employment contract. They also earn more. Even though we find no effect on entry into entrepreneurship, firms run by treated individuals nonetheless seem better managed.

We develop an empirical strategy for identifying how differences in host firms matter for interns. This methodology can be used as starting point for evaluating the increasing number of interventions where heterogeneous treatment is part of the experimental design – e.g, pairing individuals with mentors, firms with consultants, or students with teachers. In our specific case, the propensity scores for matching interns to firms is estimated by exploiting three features of our intervention design: random assignment of participants into small batches for matching purposes; matching both sides with a deferred acceptance algorithm; and estimation of a generative Bayesian model of rank formation to predict rankings. We illustrate our methodology by estimating the effect of being matched to a firm with a high-management score. We find that interns assigned to a high-management firm are more likely to run their own business in lieu of having a wage job. We also find that the assignment mechanism used *across* subjects in the treatment group has a profound effect on average program outcomes. These results outline the role of heterogeneity and matching in understanding how individuals acquire management knowledge, as well as for the design of field experiments more generally.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When Should You Adjust Standard Errors for Clustering?," *NBER Working Paper No. 24003*.

Abdulkadiroğlu, A., J. D. Angrist, Y. Narita, and P. A. Pathak (2017): "Research design meets market design: Using centralized assignment for impact evaluation," *Econometrica*, 85(5), 1373–1432.

Abdulkadiroğlu, A., and T. Sönmez (1998): "Random Serial Dictatorship and the Core from Random Endowments in House Allocation Problems," *Econometrica*, 66(3), 689.

——— (1999): "House Allocation with Existing Tenants," *Journal of Economic Theory*, 88(2), 233–260.

Abebe, G., S. Caria, M. Fafchamps, P. Falco, S. Franklin, and S. Quinn (2018): "Anonymity or Distance? Job Search and and Labour Market Exclusion in a Growing African City," *Working paper*.

Abebe, G., S. Caria, and E. Ortiz-Opsina (2018): "The selection of talent: Experimental and Structural Evidence from Ethiopia," .

Atkin, D., A. Chaudhry, S. Chaudry, A. K. Khandelwal, and E. Verhoogen (2017): "Organizational Barriers to Technology Adoption: Evidence from Soccer-ball Producers in Pakistan," *The Quarterly Journal of Economics*, 132(3), 1101–1164.

Bandiera, O., L. Guiso, A. Prat, and R. Sadun (2015): "Matching Firms, Managers, and Incentives," *Journal of Labor Economics*, 33(3), 623–681.

Bandiera, O., S. Hansen, A. Prat, and R. Sadun (2017): "CEO Behavior and Firm Performance," *Working Paper*.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007): "Remedying education: Evidence from two randomized experiments in India," *The Quarterly Journal of Economics*, 122(3), 1235–1264.

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006): "Adaptive Linear Step-up Procedures that Control the False Discovery Rate," *Biometrika*, 93(3), 491–507.

Benmelech, E., and C. Frydman (2015): "Military CEOs," *Journal of Financial Economics*, 117(1), 43–59.

Bertrand, M., and A. Schoar (2003): "Managing with Style: The Effect of Managers on Firm Policies," *The Quarterly Journal of Economics*, 118(4), 1169–1208.

Bishop, C. M. (2006): *Pattern recognition and machine learning*. Springer.

Bloom, N., R. Lemos, R. Sadun, D. Scur, and J. Van Reenen (2014): "The New Empirical Economics of Management," *Journal of the European Economic Association*, 12(4), 835–876.

Bloom, N., H. Schweiger, and J. Van Reenen (2012): "The land that lean manufacturing forgot? Management practices in transition countries 1," *Economics of Transition*, 20(4), 593–635.

Bloom, N., and J. Van Reenen (2007): "Measuring and explaining management practices across firms and countries," *The Quarterly Journal of Economics*, 122(4), 1351–1408.

Breza, E., and A. G. Chandrasekhar (2019): "Social Networks, Reputation, and Commitment: Evidence From a Savings Monitors Experiment," *Econometrica*, 87(1), 175–216.

Brooks, W., K. Donovan, and T. R. Johnson (2018): "Mentors or teachers? Microenterprise training in Kenya," *American Economic Journal: Applied Economics*, 10(4), 196–221.

Bruhn, M., D. Karlan, and A. Schoar (2018): "The impact of consulting services on small and medium enterprises: Evidence from a randomized trial in mexico," *Journal of Political Economy*, 126(2), 635–687.

Carneiro, P., J. J. Heckman, and E. J. Vytlacil (2011): "Estimating marginal returns to education," *American Economic Review*, 101(6), 2754–81.

Casey, K., R. Glennerster, E. Miguel, and M. Voors (2018): "Skills vs Voice in Local Development," Mimeo.

DellaVigna, S., and D. Pope (2018): "Predicting experimental results: Who knows what?," *Journal of Political Economy*, forthcoming.

EFRON, B., AND R. THISTED (1976): "Estimating the Number of Unseen Species: How Many Words did Shakespeare Know?," *Biometrika*, 63(3), 435–447.

ELLISON, G., AND R. HOLDEN (2013): "A Theory of Rule Development," *The Journal of Law, Economics, & Organization*, 30(4), 649–682.

FERNANDEZ, M. A. (2018): "Deferred Acceptance and Regret-Free Truthtelling: A Characterization Result," .

FRANKLIN, S. (2018): "Location, search costs and youth unemployment: experimental evidence from transport subsidies," *The Economic Journal*, 128(614), 2353–2379.

GALE, D., AND L. S. SHAPLEY (1962): "College admissions and the stability of marriage," *The American Mathematical Monthly*, 69(1), 9–15.

GOSNELL, G. K., J. A. LIST, AND R. D. METCALFE (2019): "The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains," *NBER Working Paper No. 25620*.

GUISO, L., P. SAPIENZA, AND L. ZINGALES (2015): "The Value of Corporate Culture," *Journal of Financial Economics*, 117(1), 60–76.

HECKMAN, JAMES J.AND VYTLACIL, E. (2005): "Structural equations, treatment effects, and econometric policy evaluation 1," *Econometrica*, 73(3), 669–738.

HECKMAN, J. J., AND E. VYTLACIL (2001): "Policy-Relevant Treatment Effects," *The American Economic Review Papers & Proceedings*, 91(2), 107–111.

HECKMAN, J. J., AND E. VYTLACIL (2007): "Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. 6, pp. 4875–5143.

HSIEH, C.-T., AND P. J. KLENOW (2009): "Misallocation and Manufacturing TFP in China and India," *The Quarterly journal of economics*, 124(4), 1403–1448.

HUMPHRIES, J. E. (2017): "The Causes and Consequences of Self-Employment over the Life Cycle," Job Market Paper, University of Chicago.

JAYACHANDRAN, S., J. DE LAAT, E. F. LAMBIN, C. Y. STANTON, R. AUDY, AND N. E. THOMAS (2017): "Cash for carbon: A randomized trial of payments for ecosystem services to reduce deforestation," *Science*, 357(6348), 267–273.

KAPLAN, S. N., M. M. KLEBANOV, AND M. SORENSEN (2012): "Which CEO Characteristics and Abilities Matter?," *The Journal of Finance*, 67(3), 973–1007.

KLEPPER, S., AND S. SLEEPER (2005): "Entry by spinoffs," *Management Science*, 51(8), 1291–1306.

MALMENDIER, U., G. TATE, AND J. YAN (2011): "Overconfidence and Early-life Experiences: The Effect of Managerial Traits on Corporate Financial Policies," *The Journal of finance*, 66(5), 1687–1733.

McKenzie, D., and C. Woodruff (2017): "Business practices in small firms in developing countries," *Mangement Science*, 63(3), 2773–3145.

McVitie, D. G., and L. B. Wilson (1970): "Stable Marriage Assignment for Unequal Sets," *BIT Numerical Mathematics*, 10(3), 295–309.

——— (1971): "The stable marriage problem," *Communications of the ACM*, 14(7), 486–490.

Muendler, M.-A., and J. E. Rauch (2018): "Do Employee Spinoffs Learn Markets From Their Parents? Evidence From International Trade," *NBER Working Paper No. 24302*.

Narita, Y. (2018): "Towards an ethical experiment," Discussion paper, Cowles Foundation Discussion Paper No. 2127.

Rigol, N., R. Hussam, and B. Roth (2018): "Targeting High Ability Entrepreneurs Using Community Information: Mechanism Design in the Field," Job Market Paper.

Rosenbaum, P. R., and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.

Roth, A. E. (1984): "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *Journal of political Economy*, 92(6), 991–1016.

Roth, A. E., and M. A. O. Sotomayor (1990): *Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis*, Econometric Society Monographs. Cambridge University Press.

Rubin, D. B. (2004): *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons.

Trapp, A., A. Teytelboym, A. Martinello, T. Annderson, and N. Ahani (2018): "Placement optimization in refugee resettlement," Discussion paper, Lund Department of Economics Working Paper 2018:23.

Table 1: **Description of young professionals and host firms**

	N	Mean	SD	p25	p50	p75
Male	1,637	0.76	0.43	1	1	1
University degree	1,637	0.77	0.42	1	1	1
Years of schooling	1,576	15.76	1.75	15	16	17
Years since graduation	1,632	2.57	2.37	1	1	4
Field of study: STEM	1,626	0.58	0.49	0	1	1
Field of study: business	1,626	0.21	0.41	0	0	0
Born in Addis Ababa	1,635	0.27	0.45	0	0	1
Son/daughter of household head	1,635	0.33	0.48	0	0	1
Household head or spouse	1,635	0.37	0.48	0	0	1
Wage employed	1,637	0.25	0.43	0	0	1
Monthly wage if employed	405	3,657	2,615	2,000	2,300	4,600
Self-employed	1,637	0.07	0.26	0	0	0
Monthly profit if self-employed	86	11,914	23,622	2,000	5,000	14,000
Has searched for wage job	1,636	0.80	0.40	1	1	1
Has thought of starting a business	1,637	0.28	0.45	0	0	1

Notes: This table reports descriptive information on key demographic and employment variable from the baseline survey. All monetary variables are in nominal Ethiopian birr.

Table 2: **Main outcomes on employment**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Outcome:	Self-employed	Self-emp. hours	Profit income	Wage work	Perm. work	Managerial work	Wage work hours	Wage income
Dummy: Treated	0.00	-0.01	108.44	0.03	0.04	0.02	0.41	265.25
	(0.01)	(0.08)	(180.66)	(0.02)	(0.02)	(0.01)	(0.14)	(88.50)
	[0.72]	[0.87]	[0.55]	[0.05]*	[0.01]**	[0.15]	[0.00]***	[0.00]***
	{0.45}	{0.48}	{0.38}	{0.07}*	{0.03}**	{0.14}	{0.01}**	{0.01}**
Control mean (follow-up)	0.12	0.71	923.07	0.64	0.51	0.12	4.91	2520.80
Control mean (baseline)	0.07	0.35	438.47	0.26	0.19	0.04	1.76	853.33
Observations	3,110	3,121	3,077	3,121	3,121	3,121	3,121	3,105

Note: In this table we report the *intent-to-treat* estimates of the placement on primary employment outcomes. These are obtained by least-squares estimation of equation 1. Below each coefficient, we report a standard error in parenthesis, a p-value in brackets, and a q-value in curly braces. Standard errors allow for clustering at the level of the individual. q-values are obtained using the sharpened procedure of (Benjamini, Krieger, and Yekutieli, 2006). We denote significance using $*$ for 10%, $**$ for 5% and $***$ for 1%.

Table 3: Effects of treatment on management confidence and management practices

	(1)	(2)	(3)	(4)	(5)	(6)
	Confidence		Manangement Practices			
Outcome:	Sum	Index	Overall	Marketing	Recording	Financial
Dummy: Treated	0.23	0.04	0.08	0.07	0.11	0.05
	(0.07)	(0.01)	(0.05)	(0.06)	(0.09)	(0.06)
	[0.00]***	[0.00]***	[0.09]*	[0.22]	[0.19]	[0.47]
	{0.00}***	{0.00}***	{0.42}	{0.42}	{0.42}	{0.42}
Control mean (follow-up)	9.78	0.02	-0.02	0.00	-0.05	-0.01
Control mean (baseline)	9.61	-0.04	0.07	0.02	0.17	0.02
Observations	3,121	3,121	396	396	396	396

Note: In this table we report the *intent-to-treat* estimates of the placement on primary employment outcomes. These are obtained by least-squares estimation of equation 1. Below each coefficient, we report a standard error in parenthesis, a p-value in brackets, and a q-value in curly braces. Standard errors allow for clustering at the level of the individual. q-values are obtained using the sharpened procedure of (Benjamini, Krieger, and Yekutieli, 2006), for each pre-specified outcome family. Note that the family for outcomes in columns (1) and (2) includes the outcomes reported in Appendix Table A.13. We denote significance using $*$ for 10%, $**$ for 5% and $***$ for 1%.

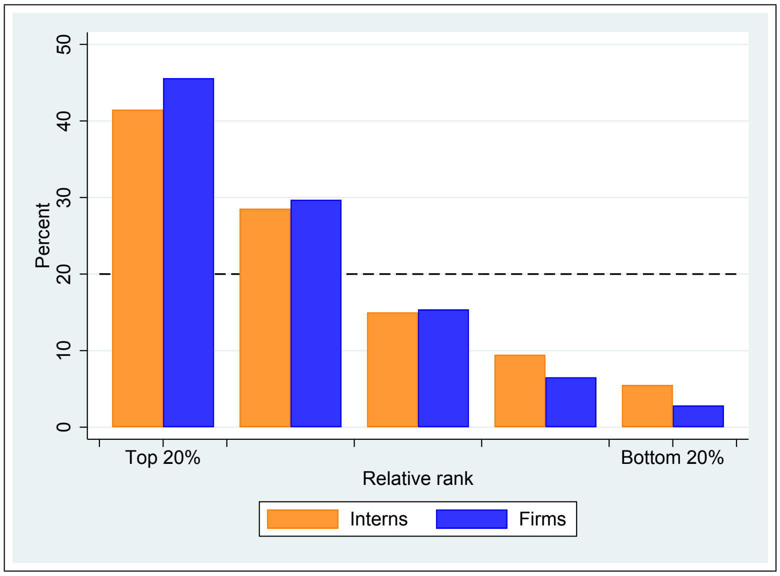
s

Table 4: **Occupation outcomes and host management quality: Baseline balance**

Outcome:	(1) Self-employed	(2) Self-emp. hours	(3) Profit income	(4) Wage work	(5) Perm. work	(6) Managerial work	(7) Wage work hours	(8) Wage income	(9) Total income
A. Baseline balance regression without controls									
Dummy: High management	0.0235	0.253	188.6	0.0705**	0.0534*	-0.00656	0.519**	267.1**	446.1
	(0.0205)	(0.161)	(239.4)	(0.0321)	(0.0290)	(0.0146)	(0.238)	(133.6)	(276.3)
Observations	704	704	695	704	704	704	704	700	691
B. Linear control function for propensity score									
Dummy: High management	-0.00345	0.137	70.96	0.0400	0.0441	-0.00981	0.386	43.12	112.1
	(0.0230)	(0.158)	(210.6)	(0.0397)	(0.0358)	(0.0153)	(0.290)	(153.2)	(266.8)
Simulated propensity score	0.0826**	0.354	361.7	0.0933	0.0287	0.00997	0.407	685.5***	1024.7**
	(0.0402)	(0.321)	(384.3)	(0.0662)	(0.0582)	(0.0262)	(0.471)	(264.6)	(462.2)
Observations	704	704	695	704	704	704	704	700	691
C. Ventile dummy control function for propensity score									
Dummy: High management	-0.00317	0.144	94.47	0.0440	0.0502	-0.00682	0.391	49.31	144.2
	(0.0231)	(0.159)	(220.1)	(0.0404)	(0.0363)	(0.0153)	(0.296)	(157.0)	(276.3)
Observations	704	704	695	704	704	704	704	700	691
D. Semi-parametric control function for propensity score									
Dummy: High management	-0.00288	0.147	91.35	0.0391	0.0448	-0.00953	0.368	36.93	124.7
	(0.0228)	(0.157)	(213.2)	(0.0396)	(0.0356)	(0.0150)	(0.289)	(153.4)	(269.4)
Observations	704	704	695	704	704	704	704	700	691
Härdle-Mammen test (p)	0.89	0.56	0.49	0.21	0.06	0.65	0.04	0.56	0.94
E. Inverse probability weighting with the propensity score									
Dummy: High management	0.00474	0.149	92.44	0.0460	0.0425	-0.00722	0.401	110.7	199.3
	(0.0204)	(0.145)	(185.6)	(0.0402)	(0.0336)	(0.0126)	(0.272)	(145.0)	(244.1)
Observations	704	704	695	704	704	704	704	700	691

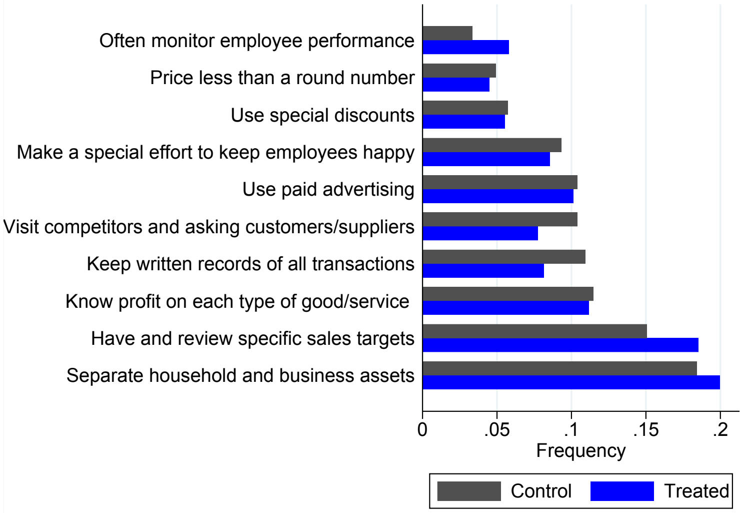
Note: In this table we report that conditioning on the simulated propensity scores balances the treatment group sample on baseline outcomes. In Panel A, we do not condition on the propensity score, and interns assigned to a high-management firm are significantly different on most baseline employment outcomes. In Panels B-E, we condition on the propensity score in different ways: using increasingly flexible control functions; and by re-weighting the observations by the inverse assignment probability (propensity score). A linear control function seems sufficiently flexible and achieves good baseline balance. We denote significance using $*$ for 10%, $**$ for 5% and $***$ for 1%.

Figure 1: **Deferred Acceptance algorithm: Summary of assignment**



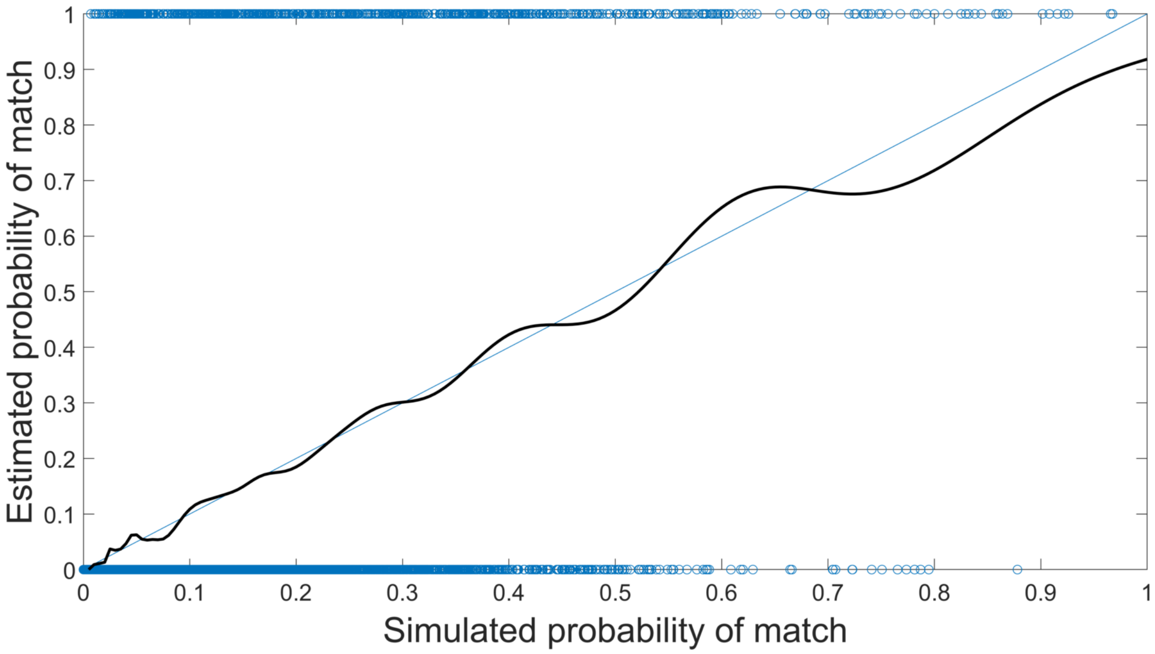
Note: This figure reports the performance of the firm-proposing DA algorithm in terms of allocating preferred counterparts to interns and firms. We report, for each quintile of the normalised rank, the fraction of firms and interns that are assigned a counterpart they rank within this quintile. Random assignment would result in bins of 20% each on average, denoted by the dashed line.

Figure 2: **Perception of most important management practice**



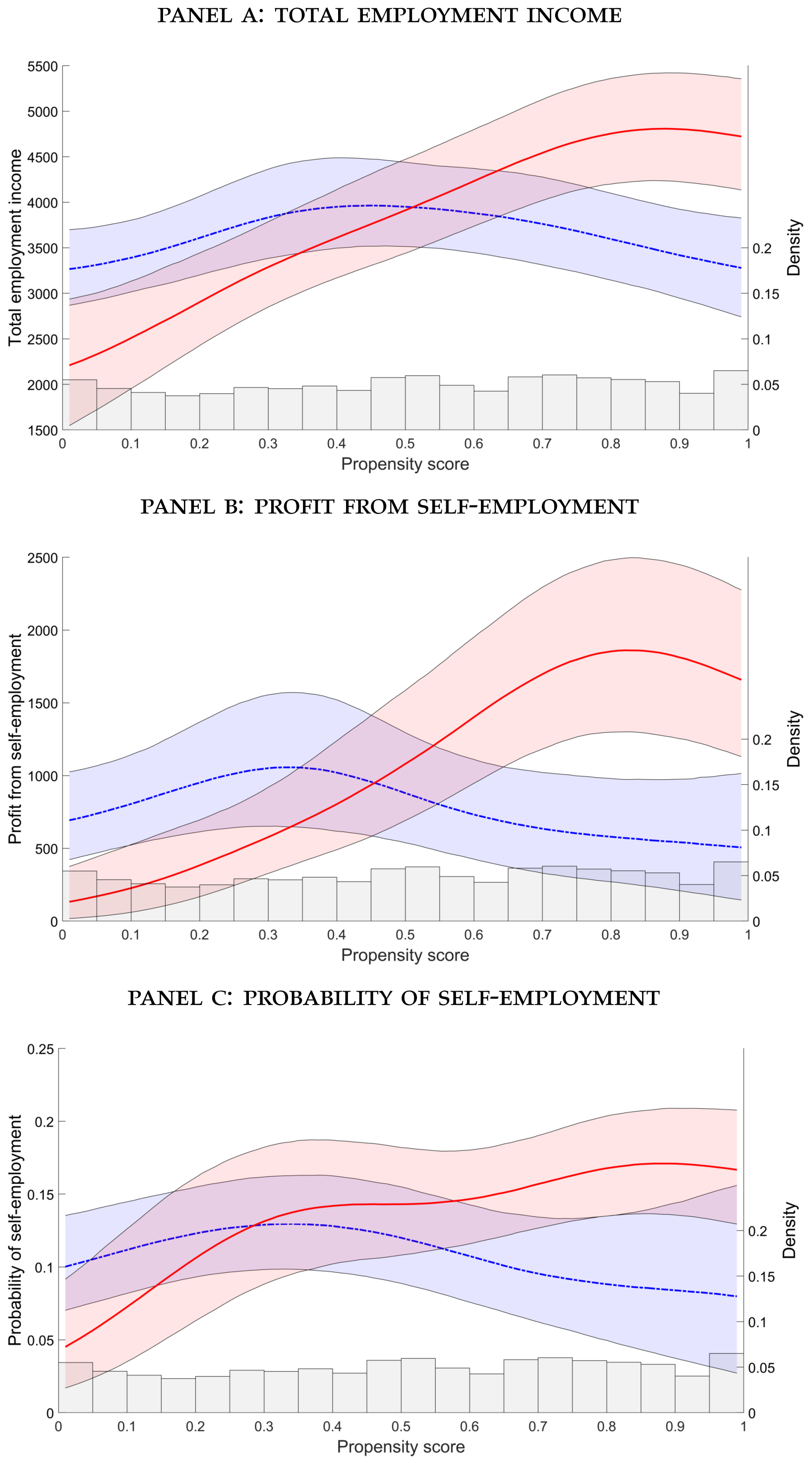
Note: This figure reports the distribution of which management practice interns perceive as the most important one, by treatment status. A Kolmogorov-Smirnoff test rejects distributional equality between treated and control with $p = 0.06$.

Figure 3: **Estimated propensity scores: A posterior predictive check**



Note: The scatterplot in this figure graphs q_{if} (the simulated assignment probability of intern i to firm f in their batch) on the x-axis and a dummy whether such assignment actually occurred m_{if} on the y-axis. The smooth and thick black line is a local linear Kernel regression with an adaptive bandwidth of 0.075; this is an estimate of $E(m_{if}|q_{if})$. In theory this should equal a 45 degree line: $E(m_{if}|q_{if})=q_{if}$ which is graphed as a thin blue line. The Kernel plot lies closely around that line.

Figure 4: **Marginal Treatment Effects under matching**



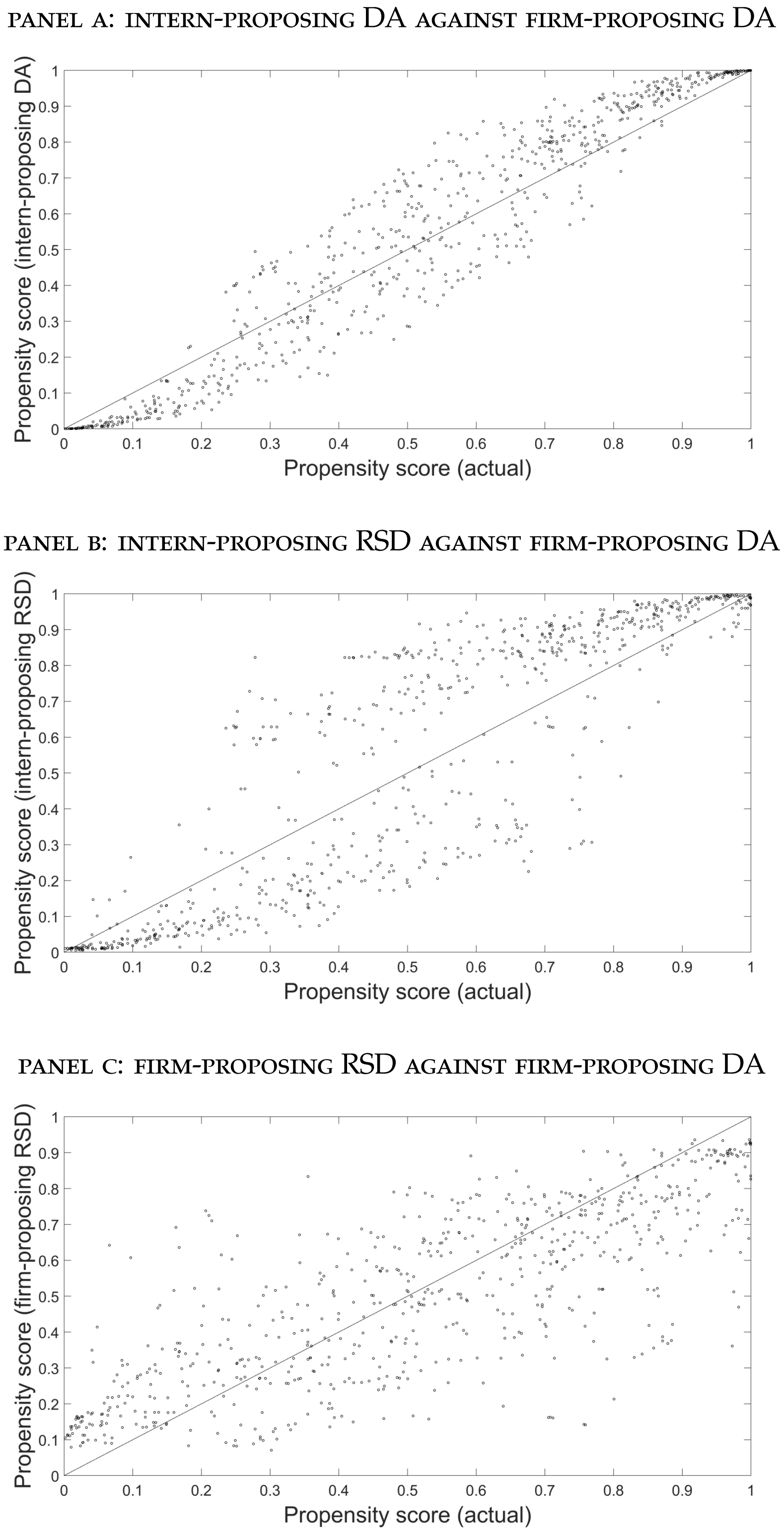
PANEL A: TOTAL EMPLOYMENT INCOME

PANEL B: PROFIT FROM SELF-EMPLOYMENT

PANEL C: PROBABILITY OF SELF-EMPLOYMENT

Note: This figure graphs the marginal treatment effects (MTE) as a function of the propensity score p of the implemented assignment mechanism. Outcomes are noted above each figure; income and profit are in monthly ETB, self-employment probability is a dummy for being self-employed. The scale corresponds to the left y-axis. The red solid curve graphs the outcome for interns assigned to a high-management firm ($y_1(p)$), the blue dashed curved graphs the outcome for interns assigned to a low-management firm ($y_0(p)$). The curves are obtained from a Kernel regression with a Gaussian Kernel and a bandwidth of 0.15. The difference between these curves is the integral of MTE over a small interval around p. Shaded areas around the curves are 90% confidence intervals. These take into account parameter uncertainty that underlies the simulated propensity scores by repeatedly drawing from the posterior distributions to obtain a posterior distribution of propensity scores. At the bottom of the graph is the histogram of propensity scores, in 20 equal-width bins (densities scale on the right y-axis).

Figure 5: **Bivariate distributions of propensity scores: Alternative mechanisms**



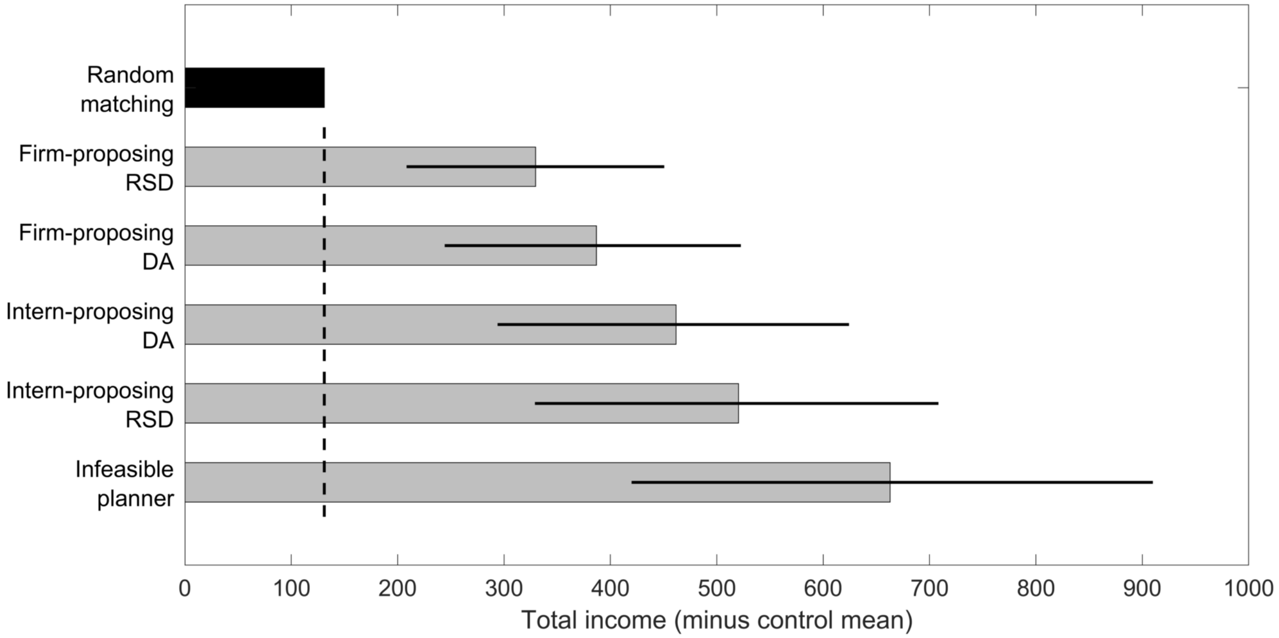
Panel A: intern-proposing DA against firm-proposing DA

Panel B: intern-proposing RSD against firm-proposing DA

Panel C: firm-proposing RSD against firm-proposing DA

Note: This figure graphs the mean posterior of propensity scores of the mechanism actually implemented (firm-proposing deferred acceptance) against alternative mechanisms, denoted in the panel titles. The figures show that intern-proposing mechanisms have a larger concentration of assignment probabilities, whereas the firm-proposing random serial dictatorship (RSD) has more dispersion.

Figure 6: **Estimated income under counterfactual mechanisms**



Note: This figure graphs the counterfactual total monthly labor income under alternative assignment mechanisms, minus the control group mean. This is based on equation 14 and the mean simulated propensity scores under different mechanisms. The black whiskers show 90% confidence intervals obtained from the posterior distribution of propensity scores.