Can Business Owners Form Accurate Counterfactuals?

Eliciting Treatment and Control Beliefs about Their Outcomes in the Alternative Treatment Status[#]

David McKenzie, *World Bank*

Abstract

A survey of participants in a large-scale business plan competition experiment, in which winners received an average of US$50,000 each, is used to elicit ex post beliefs about what the outcomes would have been in the alternative treatment status. Participants are asked the percent chance they would be operating a firm, and the number of employees and monthly sales they would have, had their treatment status been reversed. The study finds the control group to have reasonably accurate expectations of the large treatment effect they would experience on the likelihood of operating a firm, although this may reflect the treatment effect being close to an upper bound. The control group dramatically overestimates how much winning would help them grow the size of their firm. The treatment group overestimates how much winning helps their chance of their business surviving, and also overestimates how much winning helps them grow their firms. In addition, these counterfactual expectations appear unable to generate accurate relative rankings of which groups of participants benefit most from treatment.

Keywords: counterfactual elicitation, subjective expectations, randomized experiment, business growth.

JEL codes: O12, D22, C80, C31, C93

[#] I thank DFiD, the World Bank, the WLSME trust fund, and the John Templeton Foundation for funding the surveys used for this study, the Strategic Research Program (SRP) for funding work on this paper, and Martin Ravallion, the editor, associate editor, and four anonymous referees for helpful comments.

1. Introduction

I assess the extent to which business owners can form accurate counterfactuals about what would have happened to their business had an alternative business investment process taken place. I do this in the context of a large business plan competition carried out in Nigeria, where the winners received an average of US$50,000 each as business grants. A large subset of the winners was chosen randomly from a pool of semi-finalists, enabling causal impacts of the program's effects. An experimental evaluation of this program undertaken in McKenzie (2015) found that winning this competition resulted in large increases in business start-up, survival, employment growth, and sales growth. Three years after the random selection, I elicited ex post beliefs from winners in the treatment group about the counterfactual of what would have happened to their business should they have lost, and non-winners in the control group were asked similar questions about what would have happened should they have won.

Assessing the accuracy of these elicited counterfactuals is useful for at least two reasons. First, from a program evaluation standpoint it is often difficult to accurately estimate the counterfactual, especially for programs not implemented using a randomized experiment. If applicants are able to provide accurate estimates of their counterfactuals under the alternative treatment status, this would provide a means of estimating program impacts even for programs in which no suitable control group can be found.[1] Such an approach is used in a recent evaluation of workfare by Murgai et al. (2016). Second, from the standpoint of business economics, an important question is the extent to which firms can learn the optimal use of inputs and business practices through a process of trial and error. Often firms face so many other shocks that affect their profits and sales that economists struggle to have sufficient power to estimate the average effects of interventions in randomized experiments (e.g. McKenzie, 2011). Recent work by Lewis

[1] See also Ravallion (2014) for an alternative attempt to conduct "shoestring" evaluations by asking program participants in both the treated and control groups retrospective questions on how outcomes have changed, which he finds does not yield accurate measurement of the treatment effect.

and Rao (2015) has highlighted that, in the case of advertising expenditures, firms seldom have power to measure the returns of such expenditure even with millions of consumers to experiment on. The intervention studied here involves a large change to the business, and so represents a case where we might expect firms to better be able to assess the consequences of it taking place or not.

The results show that firms do not provide accurate counterfactuals even in this case of a large intervention. I find that both the control and treatment groups systematically overestimate how important winning the program would be for firm growth: the control group thinks they would grow more had they won than the treatment group actually grew, while the treatment group thinks they would grow less had they lost than the control group actually grew. This systematic bias towards thinking programs will help may help explain why firms apply for a number of programs such as business training that struggle to demonstrate impacts (McKenzie and Woodruff, 2014) and suggests that the use of elicited counterfactuals is unlikely to be a good substitute for alternative forms of impact evaluation in dynamic settings. Moreover, I do not find these counterfactuals to provide accurate relative rankings of which groups benefit most from treatment, suggesting these elicited counterfactuals are also unlikely to provide a useful means of targeting programs.

The remainder of the paper is structured as follows: Section 2 discusses related literature on eliciting counterfactuals from program participants. Section 3 describes the intervention and the process of measuring counterfactuals. Section 4 provides the results, and Section 5 concludes.

2. Related Literature

Economists have increasingly used questions on subjective expectations in a wide variety of contexts, with Manski (2004) and Delavande et al. (2011) providing reviews for work in developed and developing countries respectively. In the specific context of understanding the impacts of policies using impact evaluations, there are two key strands of literature, differing in whether expectations are assessed before or after the intervention has taken place.

The first branch of the literature aims to assess the accuracy of ex ante expectations of policy makers and program participants as to the likely effects of an intervention. An early attempt in this literature was to ask caseworkers to predict which individuals applying to a training program would have the best outcomes if they ended up getting selected for treatment. Bell and Orr (2002) find that caseworkers are able to predict who will have higher employment outcomes after training. But predicting levels is different from predicting treatment effects, and the caseworkers were not asked to predict who would benefit most from the program. Using non-experimental methods, Lechner and Smith (2007) find caseworkers do no better at allocation than random assignment would in terms of maximizing treatment impacts.

More recent studies explicitly ask expectations of treatment effects. Hirshleifer et al. (2016) ask applicants to a vocational training program and policy makers involved in their program their expectations about the likely impact of training on future employment outcomes. Comparing these expectations to the estimated treatment effects, they show that both participants and policy makers greatly overestimate the impact of the program. Likewise, Andrade et al. (2016) find government officials overestimate the success of programs designed to formalize small firms, and Groh et al. (2016) find policy makers and academic audiences overestimate the likely effects of soft skills training. These studies therefore tend to show that expectations tend to be systematically biased in favor of the intervention before it has taken place.

This paper is most directly related to a small literature which instead considers ex post expectations of counterfactuals. Smith et al. (2011) consider participants in a job training program in the U.S. that was subject to an experimental design. Recipients of the program were then asked "Do you think that the training or other assistance that you got from the program helped you get a job or perform better on the job?" They compare these self-reports of the program's effectiveness to the experimental estimates and find very little relationship between what participants say and these experimental estimates. This is an interesting finding, although as the authors acknowledge, there are a number of problems with the question asked to assess beliefs about counterfactuals: the question they use does not explicitly prompt respondents to think in counterfactual terms, does not specify a time period to assess impact over, and does not provide quantitative estimates of the counterfactual. They note that these are common problems of almost all participant evaluation studies, and recommend that future studies use more sophisticated subjective expectations questions that explicitly ask about counterfactuals. This is the approach we test in this paper, with the added advantage of also asking the control group counterfactual questions about their outcomes if they had been treated.

Murgai et al. (2016) ask participants in a workfare program in India whether they would have been employed or not, and how much they would have earned without workfare. They use this information to form the counterfactual in the context of a non-experimental evaluation. While they do not have an experimental control group to compare the treatment group's expectations to, they show that the elicited expectations of casual labor earnings are very similar to those earned by other people doing casual labor in these villages. In such a setting, in which the counterfactual to treatment is just to be doing something that the individual has been doing for years before, with little dynamics, expectations are likely to not differ very much from simply reporting past earnings, and may be reasonably accurate. McKenzie et al. (2013) provide an additional example of this, showing that Tongan immigrants in New Zealand have reasonably accurate expectations of what they would have earned had they not migrated, because this would be very similar to what they earned before migrating. In contrast, they find the control group underestimates how much they could earn by migrating.

Our setting is one which is dynamic, with firm owners attempting to grow their businesses even in the absence of taking up the program. On top of this general growth, business incomes are highly volatile, and many idiosyncratic factors can determine whether firms survive, and how much they can expand. As a result, with business start-ups, forming expectations requires more than simply saying that this year will be like all the years before unless the program is taken up.

3. Intervention and Measurement

This section begins by briefly summarizing the key details of the intervention that individuals will be asked to assess the counterfactual for, before then discussing the survey questions used to measure these counterfactuals.

3.1 The YouWiN! Business Plan Competition

The Youth Enterprise With Innovation in Nigeria (YouWiN!) program is a nationwide business plan competition for young entrepreneurs in Nigeria that was launched in late 2011. Applicants applied online, and a first screening was used to invite the top 6,000 out of almost 24,000 entrants to attend a four-day business plan training course. These individuals then submitted business plans, which were scored, with the top 2,400 being selected as semi-finalists. Among the semi-finalists, those with the highest overall scores were selected as winners with probability one, and then a random draw was used to select the remaining 729 winners from a group of 1,841 semi-finalists. Winners received a grant which averaged approximately US$50,000 each, which was paid out over a period of a year in four tranches. Three follow-up surveys track these firms with the last survey taking place between September 2014 and February 2015. This corresponds to three years after application, and between 12 and 18 months after firms had received their last tranche payment from the program.

Applicants could apply as either a *new firm* with an idea for starting a business, or as *an existing firm* with the goal of expanding their business. Random selection was stratified by this firm type, as well as by geographic region and gender. The typical applicant in this experimental pool is aged around 30, has university education, and owns a computer. As such, concerns about the ability of individuals to think about probabilities and alternative outcomes are likely to be less than with less-educated respondents who are the focus of studies of smaller firms in developing countries. Only 17.4 percent of the sample is female, reflecting that males were much more likely to apply for the program.

In McKenzie (2015), I use this randomized assignment and follow-up surveys to measure the impacts of this program. This is done separately for the new firm applicants and existing firm applicants, reflecting their different starting statuses. Winning is found to have large treatment impacts on both groups of firms. By the third follow-up survey, new firm applicants are 37.3 percentage points more likely to be operating a firm, have 5.2 more workers per firm, and are 22.9 percentage points more likely to have a firm of 10 or more employees. The impact on sales in the third round is positive at 64,500 naira per month ($410), but not statistically significant, although an aggregate index of sales and profits shows a significant increase. By this last survey round, existing firm applicants are 19.6 percentage points more likely to be operating a firm, have 4.4 more workers per firm, and are 20.6 percentage points more likely to have reached 10 or more workers. Winning also significantly increases sales by 338,000 naira per month (US$2153), which is a 66 percent increase on the control mean.

This setting has several favorable characteristics for considering the accuracy of counterfactuals. It involves a large treatment effect, and a relatively educated set of individuals. The intervention is a relatively simple one – the main benefit of winning being the $50,000 grant winners receive – so that it should be reasonably easy for the control group to also understand what exactly the treatment is. This contrasts with job or business training programs, where the control group is unlikely to have accurate knowledge of what the exact content of the intervention is, making it harder for them to think about likely impacts.

3.2 Eliciting Business Owner Counterfactuals

The last follow-up survey elicited counterfactuals for business ownership, employment, and sales for both groups. Applicants in the control group were asked the following questions:

Imagine you had won the YouWin! Competition in the first round. I would like you to tell me what you think you would be doing right now in terms of running a business or working.

*a) What do you think is the percent chance (out of 100) that you would be running a business today if you had won YouWin!?*____________________

b) If you had won YouWin!, and were running a business today, how many employees in total do you think you would have today?

c) If you had won YouWin!, and were running a business today, how much do you think your sales would have been for the last month?

Winning applicants in the treatment group were asked:

Imagine you had NOT won the YouWin! Competition in the first round. I would like you to tell me what you think you would be doing right now in terms of running a business or working.

*a) What do you think is the percent chance (out of 100) that you would be running a business today if you had NOT won YouWin!?*____________________

b) If you had NOT won YouWin!, and were running a business today, how many employees in total do you think you would have today?

c) If you had NOT won YouWin!, and were running a business today, how much do you think your sales would have been for the last month?

Question a) in both cases is a probabilistic expectation following the percent chance formulation of Dominitz and Manski (1997) and Manski (2004), which has been used with tertiary-educated individuals in a developing country context by McKenzie et al. (2013). In contrast b) and c) are non-probabilistic expectations questions, following the approach of directly asking what you expect used by Jensen (2009) among others. A potential critique of such questions is that it is unclear whether respondents are answering with a mean, median, mode, or something else. The reason for using such a method was to enable surveys to be done by telephone or online, in addition to face-to-face.[2] Delavande et al. (2011) provide evidence that this form of direct expectations question is strongly correlated with the mean of a subjective probability distribution, but can be more susceptible to outliers. As a result, I examine the responses for large outliers and trim one extreme outlier (noted below).

3.3 Assessing the Accuracy of These Counterfactuals

Let Y_i be the realized outcome in the third follow-up survey for individual i, and E_i be the counterfactual they express when answering the questions described above. Let T denote the treatment group and C the control group. Then the mean expected treatment effect of changing treatment status for group $g \epsilon \{T, C\}$ is given by:

$$\beta_g = \frac{1}{N_g} \sum_{i \epsilon g} (E_i - Y_i) \qquad (1)$$

where N_g denotes the number of individuals in group g. Note that this will be negative if individuals expect to be worse off by changing treatment status, which is the case for the treated group. However, to enable comparability to the estimated treatment effect I report -β_T. This is estimated separately for the new firm applicants and existing firm applicants. The first approach to assessing the accuracy of the counterfactuals expressed by applicants is then to compare this estimate of the treatment effect to experimental treatment effects reported in McKenzie (2015).

For the continuous outcomes of employment and sales, I also see how accurate the control group is in assessing what their outcomes would look like if they had won by comparing the distribution of the counterfactual outcome E_i for the control group to the distribution of realized outcomes Y_i for the

[2] Surveys were typically asked face-to-face using tablets programmed with the questionnaire. However travel restrictions in some parts of Nigeria due to the security situation meant some individuals were given phone or online surveys, with these alternative methods also used to chase up individuals who were not found on multiple attempts to interview them in person.

treatment group. Since treatment assignment was randomized, the treatment group distribution should provide a consistent estimate of the true counterfactual for the distribution of outcomes of the control group had they been treated.[3] Likewise one can see how accurate the treatment group is in assessing what the outcomes would have been had they not won by comparing the distribution of counterfactual outcomes for the treated group to the distribution of realized outcomes for the control group.

While the purpose of this paper is to assess whether program participants can provide reasonable estimates of counterfactuals that could be used to measure the impacts of a program, it is useful to note two related issues around the use of such measures. The first is that being able to provide an estimate of the counterfactual in a survey may be a distinct cognitive task from being able to consider counterfactuals when making decisions. Just as in Milton Friedman's famous pool player analogy, people may be able to take decisions that incorporate counterfactual expectations even if they cannot explain those expectations. Comparing the expectations of those who chose to participate in a program versus those who did not might offer insight into this point in future work, by enabling one to test whether people self-select into programs on the basis of their expected treatment effects, as in Heckman et al. (2006). Nevertheless, irrespective of whether people behave as if they know counterfactuals, if they are unable to express these as survey answers, then eliciting counterfactuals will not yield good estimates of the treatment effect.

Secondly, there is a growing literature that uses subjective beliefs about counterfactuals to estimate choice models (e.g. Lochner 2007; Delavande 2008; Arcidiacono et al. 2012). Such models can still be used even if beliefs are inaccurate, so long as individuals are able to report their true beliefs. But knowing

[3] This would not be the case if there was systematic difference in survey attrition between the two groups. The survey response rate was 83%, and McKenzie (2015) shows that there does not appear to have been a bias from this attrition.

whether or not these counterfactuals are accurate will still be important for assessing welfare in such models and for understanding why individuals might make suboptimal choices.

3.4 Expectations of the Accuracy of Ex-Post Counterfactuals

Ex ante one can think of several possibilities as to the likely accuracy of the counterfactuals. One possibility might be that business owners are not systematically wrong about the impact of the program, so that the average treatment impact estimated using the counterfactuals should be similar to the experimental treatment effect. One potential reason to think this is that in applying for the competition the business owners had spent four days learning how to develop a business plan, and then had to submit a detailed business plan outlining how they would use the grant to develop their business. The control group have therefore all had to previously make projections and plans for business growth based on what would happen if they won, so that we are asking about a counterfactual they have spent time thinking about.

A second possibility is that behavioral factors lead to systematic biases in how individuals think of these counterfactuals and how they respond to survey questions about them. One such issue is cognitive dissonance, in which individuals may report (and even believe) answers that are consistent with their behavior and past actions (e.g. Bertrand and Mullainathan, 2001). For example, the treatment group may wish to attribute their success to their own hard work and talent rather than to winning the program, in which case they would underestimate the program effect. Conversely they may fail to take account of the progress they would have made anyway, attributing all their growth to the program and overstating the effect. The control group might want to make themselves feel better about missing out on the program by understating its impact (therefore saying to themselves that not winning does not matter that much). Conversely they may want to make themselves feel better about their current level of business success by overstating the impact of the program (saying to themselves I may be small today, but it is only because I did not win and if I had that grant I would be very successful).

A related behavioral issue may be the endowment effect (Thaler, 1980), whereby people's value things more once they own them. In this context, it may lead entrepreneurs to value the grant more because they received them, so to give a higher value to the effect of winning the grant than the control group does.

One potential way to assess the likelihood of cognitive dissonance and endowment effects biasing reports is to measure beliefs before and after a decision or intervention, and see the extent to which people update. This is not possible here, since beliefs were not elicited at the time of applying for the program. However, Zafar (2011) does such an exercise in the context of college major choice, finding that cognitive dissonance does not appear to be an issue in that context.

A third possibility is that individuals in the general population have correct expectations about the impact of the program, but that some people are overly optimistic and others overly pessimistic. Following Heckman et al. (2006) we would then expect the pool of applicants to be overly optimistic. If applicants then do not update their expectations after having experienced the program, the elicited counterfactuals for program participants will be over-optimistic about the effects of the intervention.

As such there are plausible reasons why counterfactuals may be accurate, but also plausible reasons why they could be biased in either direction.

4. Results

4.1 Counterfactuals about the Likelihood of Operating a Business

The top panel of Figure 1 plots the distribution of responses to the percent chance that control group individuals who applied as new firms would be running a business had they won the business plan competition. Vertical lines then indicate the mean counterfactual they express, along with the observed mean in the treatment group. The bottom panel similarly plots the distribution of counterfactual responses for treatment group individuals who applied as new firms in terms of their beliefs about the likelihood they would be running a business had they lost. Vertical lines indicate their mean counterfactual, along with the observed rate of running a business in the control group. Figure 2 does the same but for existing firm applicants.

Among the control group, 50.4 percent of new firm applicants and 76.3 percent of existing firm applicants are running businesses (Table 1, column 2). We see from the top panels of Figures 1 and 2 that the modal response is for them to believe there is a 100 percent chance they would be running a business had they won. The third column of Table 1 shows the mean percent chance of operating a business is 90 percent in both cases. Using equation (1), the estimated counterfactual treatment effect is thus 90.6-50.4 = 40.2 percentage points for new firm applicants, and 14.1 percentage points for existing firms. The last column of Table 1 provides the experimental treatment effects, which are 37.3 percentage points for new firms and 19.6 percentage points for existing firms.

In both cases the treatment effect using the counterfactuals lies within the relatively narrow confidence intervals of the experimental treatment effect. So in this case the control group is providing a reasonably accurate counterfactual. However, note that the treatment group has very high business ownership rates (96 and 97 percent as per column 2), making the upper bound of 100 percent not that over-optimistic. As a result, there is relatively little room for applicants to overstate the treatment effect here.

The bottom panels of Figures 1 and 2 show much wider dispersion in the responses of the treatment group to the percent chance that they would run a business had they not won the business plan competition. For both new and existing applicants the mode is at 50 percent, with approximately 17 percent of firms in both cases giving this response. Heaping at 50 percent is sometimes interpreted to reflect uncertainty (de Bruin et al. 2000). Asking a direct follow-up question of respondents as to whether a 50 percent response simply indicates uncertainty has sometimes been used to better understand this response. This was not done in this study. However, the fact that we do not see this same heaping at 50 percent for the control group's responses suggests that in our case the mass at 50 percent may largely be genuine (although reflecting some rounding). Moreover, I show robustness to reallocating some of the 50 percent responses below.

Applying the same approach as for the control group, the bottom two rows of Table 1 show that the treatment effect estimated using the counterfactual responses is 45.2 percent for new applicants in the treated group and 41.5 percent for existing applicants. The estimates for the existing businesses are overstated by a large amount. They appear to be overestimating how much the grant has helped their firm to survive, with their mean expected chance of operating a business of 55 percent more than 20 percentage points below the rate at which the control group existing firms are actually operating. To examine the sensitivity of this result to the heaping at 50 percent, as a robustness check I randomly remove half of the existing firm treated group that report 50 percent. This only raises the mean percent chance of running a business to 55.6 percent from the 55.1 percent reported in Table 1. Hence the heaping at 50 percent is not responsible for the large overestimation of the treatment effect by treated existing firms.

4.2 Counterfactuals about the Size of the Firm: Number of Employees and Sales

Our counterfactual question on employment asks about the number of employees conditional on being in business. Column 2 of Table 2 shows that the control group averages 6.9 (new firms) and 7.6 (existing firms) employees, while the treated firms are larger at 9.7 (new firms) and 10.2 (existing firms) employees. Columns 3, 4, and 5 of this table provide the mean employees given in the counterfactual, a trimmed mean which drops the top 5 percent to alleviate the concerns previously mentioned about the format of this question being susceptible to outliers, and the median response. The last two columns then contain the estimated treatment effect using the counterfactual and the experimental treatment effect respectively.[4]

Consider the control group firms. On average they expect that they would have 23 employees had they won the business plan competition, which is more than twice the actual mean level reported by the treated firms. Trimming the mean or using the median still results in a large overstatement. As a result, the estimated average treatment effect of 16 workers using this counterfactual is much larger than the experimental treatment effects of 2.2 to 2.5 workers. Figure 3 examines this overstatement of the treatment effect across the distribution by comparing the distribution of counterfactual employment reported by the control group to the actual distribution of employment for the treatment group. We see the counterfactual distribution lies to the right of the actual distribution at all quantiles, confirming that the control group overestimates the size their firm would be if they had won.

The treated group firms in contrast underestimate how many employees they would have if they had not won, and therefore also overstate the impact of the program. The estimated impact using the counterfactuals they report is 5.9 to 6.5 workers. While outside the confidence intervals for the experimental treatment effects, they are much closer than the control group were. Figure 4 confirms this. We see the general tendency across quantiles for the treated group to underestimate what their employment would be if they did not win, but that the gap between their counterfactual distribution and the control distribution is not as large as was the case in Figure 3. The greater accuracy about counterfactuals of the treated group compared to the control group may reflect that it is easier for them to picture what business would be like without the grant, since they were in this position three years

[4] Note that I use here the difference in treatment and control employment conditional on being in business, reported in the appendix of McKenzie (2015). These estimates correspond most closely to what the expectations questions ask. Alternatively one can combine the elicited probability of operating a firm with the counterfactual number of employees conditional on operating, and then compare these to the treatment effect on unconditional employment reported in Table 3 of McKenzie (2015). This gives similar results.

earlier, whereas the control group has never experienced a situation similar to the one they are being asked about. This concurs with the discussion in section 2 that it may be easier to provide expectations about more familiar events.

The results are similar for sales. Figure 5 shows that for both new and existing firms, the control group expects that sales would be greater had they won than sales actually are for the treatment group, therefore overstating the treatment impact. Figure 6 shows that the treatment group expects sales would be lower had they lost than sales actually are for the control group, therefore overstating how much they have benefited from winning. In particular they do not expect to see the very high levels of sales reported by some of the control firms.[5]

Taken together, these results provide strong evidence that business owners in this program do not report accurate counterfactuals, but instead overestimate how important the program is for business growth. This is in spite of the program having large and significant treatment effects; it is just that firms expect even larger impacts. This result that participants in a program over-estimate the benefits of it is consistent with work showing that the ex ante expectations of policy makers developing programs and individuals participating in them also tend to be overstated (e.g. Hirshleifer et al, 2016; Groh et al, 2016).

4.3 Relative Ranks and Heterogeneity

Even if the expectations are inaccurate in terms of levels, are they useful in terms of telling us which groups will benefit most from treatment? To examine this question, I take the group with the highest sample size – the control group new applicants – and examine heterogeneity in their employment counterfactual expectations by some key observed characteristics: gender, education, ability (measured

[5] One explanation of this would be if all the treatment firms had an identical distribution of potential sales. Then since they are reporting the mean of the distribution or something close to it, this deviation at the top represents some control firms getting lucky draws from this distribution, which exceed the mean. But if this were the case, we would expect that other control firms would get unlucky draws that are less than the mean, and so that the two distributions should cross around the mean. We do not see this crossing in practice.

by digitspan recall), and location of business. Table 3 reports the resulting treatment effects according to these expectations, and compares them to the estimated experimental treatment effects.

Unfortunately for the purpose of this exercise, in this context there is not statistically significant heterogeneity in treatment effects according to any of these observed differences. We see two key results in Table 3. First, every single group dramatically overestimates what the employment impact of winning the competition would be – so it is not the case that expectations are accurate for one gender and not for another, or accurate for university-educated participants and for less educated participants, etc. Indeed I cannot reject equality of counterfactual treatment effects by gender (p=0.42), education (p=0.98), digitspan (p=0.79) or location (p=0.92). Second, the rankings of treatment effects according to counterfactuals are inaccurate with respect to the rankings according to the experimental treatment effects. For example, males expect a higher treatment effect than females, but if anything the female effect is larger; those in the cities of Abuja and Lagos expect higher treatment effects, but if anything have lower treatment effects. While limited, this evidence casts doubt on the ability of this method to target interventions at who benefits most.

5. Conclusions

I find that participants in a business plan competition do not provide accurate counterfactuals when asked what would have happened had they had the alternative treatment status. Instead both the treatment and control groups systematically overestimate how important winning the competition is or would be for the growth of their businesses. In addition, the counterfactuals are also inaccurate for generating relative rankings of which types of businesses benefit most from treatment.

Although these results only apply to one experiment, they are consistent with a broader body of literature showing that policy makers and program participants tend to have over-optimistic ex ante expectations about how effective different policy interventions are likely to be. The difference here is that these beliefs are not corrected even after having experienced the program. The result is that eliciting counterfactuals directly from program participants does not provide an accurate way of assessing the impact of this program.

This exercise also suggests several directions for further research. The first is replication in other experimental settings, including ones in which the intervention has less impact or where the counterfactual state is not so dissimilar from circumstances the individual has experienced before. Secondly, eliciting subjective expectations of program impacts for both participants and non-participants prior to the program beginning, and then following up with ex-post counterfactual questions of the type set out here will enable a deeper understanding of why bias may occur. Thirdly, there seems scope for experimenting with alternative elicitation methods and question wordings, including approaches which ask about impacts for the average program participant, as well as their personal counterfactual.

References

Andrade, Gustavo Henrique de, Miriam Bruhn, and David McKenzie (2016) "A Helping Hand or the Long Arm of the Law? Experimental Evidence on What Governments Can Do to Formalize Firms", *World Bank Economic Review*, 30(1): 24-54.

Arcidiacono, Peter, V. Joseph Hotz and Songman Kang (2012) "Modeling College Major Choices Using Elicited Measures of Expectations and Counterfactuals", *Journal of Econometrics* 166(1): 3-16.

Bell, Stephen and Larry Orr (2002) "Screening (and Creaming?) Applicants to Job Training Programs: the AFDC Homemaker-Home Health Aide Demonstrations", *Labour Economics* 9(2): 279-301.

Bertrand, Marianne and Sendhil Mullainathan (2001) "Do People Mean What They Say? Implications for Subjective Survey Data", *American Economic Review Papers and Proceedings* 91(2): 67-72.

Bruine de Bruin, Wandi, Baruch Fischhoff, Susan G. Millstein and Bonnie L. Halpern-Felsher (2000) "Verbal and Numerical Expressions of Probability. „It‟s a Fifty-Fifty Chance‟," *Organizational Behavior and Human Decision Processes* 81: 115-31.

Delavande, Adeline (2008) "Pill, patch, or shot? Subjective expectations and birth control choice", *International Economic Review* 49(3): 999-1042.

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

Dominitz, J. and Charles Manski (1997) “Using expectations data to study subjective income expectations”, *Journal of the American Statistical Association* 92 (439), 855–867.

Groh, Matthew, Nandini Krishnan, David McKenzie and Tara Vishwanath (2016) “The impact of soft skills training on female youth employment: Evidence from a randomized experiment in Jordan”, *IZA Journal of Labor and Development* 5:9.

Heckman, James, Sergio Urzua and Edward Vytlacil (2006) “Understanding Instrumental Variables in Models with Essential Heterogeneity”, *Review of Economics and Statistics* 88(3): 389-432.

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

Jensen, Robert (2009) “The perceived returns to education and the demand for schooling”, *Quarterly Journal of Economics* 125 (2), 515–548.

Lechner, Michael and Jeffrey Smith (2007) “What is the value added by caseworkers?”, *Labour Economics* 14(2): 135-51.

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

Lochner, Lance (2007) “Individual Perceptions of the Criminal Justice System”, *American Economic Review* 97(1): 444-460.

Manski, Charles (2004) “Measuring expectations”, *Econometrica* 72 (5), 1329–1376

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

McKenzie, David (2015) “Identifying and Spurring High-Growth Entrepreneurship: Experimental Evidence from a Business Plan Competition”, *BREAD Working Paper no. 462.*

McKenzie, David, John Gibson, and Steven Stillman (2013) “A land of milk and honey with streets paved with gold: Do emigrants have over-optimistic expectations about incomes abroad?”, *Journal of Development Economics* 102: 116-127.

McKenzie, David and Christopher Woodruff (2014) “What are we learning from business training evaluations around the developing world? “, *World Bank Research Observer*, 29(1): 48-82

Murgai, Rinku, Martin Ravallion, and Dominique van de Walle (2016) “Is workfare cost effective against poverty in a poor labor-surplus economy?”, *World Bank Economic Review*, 30(3): 413-45.

Ravallion, Martin (2014) “Can we trust shoestring evaluations?”, *World Bank Economic Review* 28(3): 413-31.

Smith, Jeffrey, Alexander Whalley, and Nathaniel Wilcox (2011) “Are program participants good evaluators?”, mimeo. University of Michigan.

Thaler, Richard (1980) “Toward a positive theory of consumer choice”, *Journal of Economic Behavior & Organization* 1 (1): 39–60

Zafar, Basit (2011) “Can Subjective Expectations Data be Used in Choice Models? Evidence on Cognitive Biases”, *Journal of Applied Econometrics* 26: 520-44.

Figure 1: Counterfactual Beliefs of Likelihood of Operating a Firm for New Firm Applicants

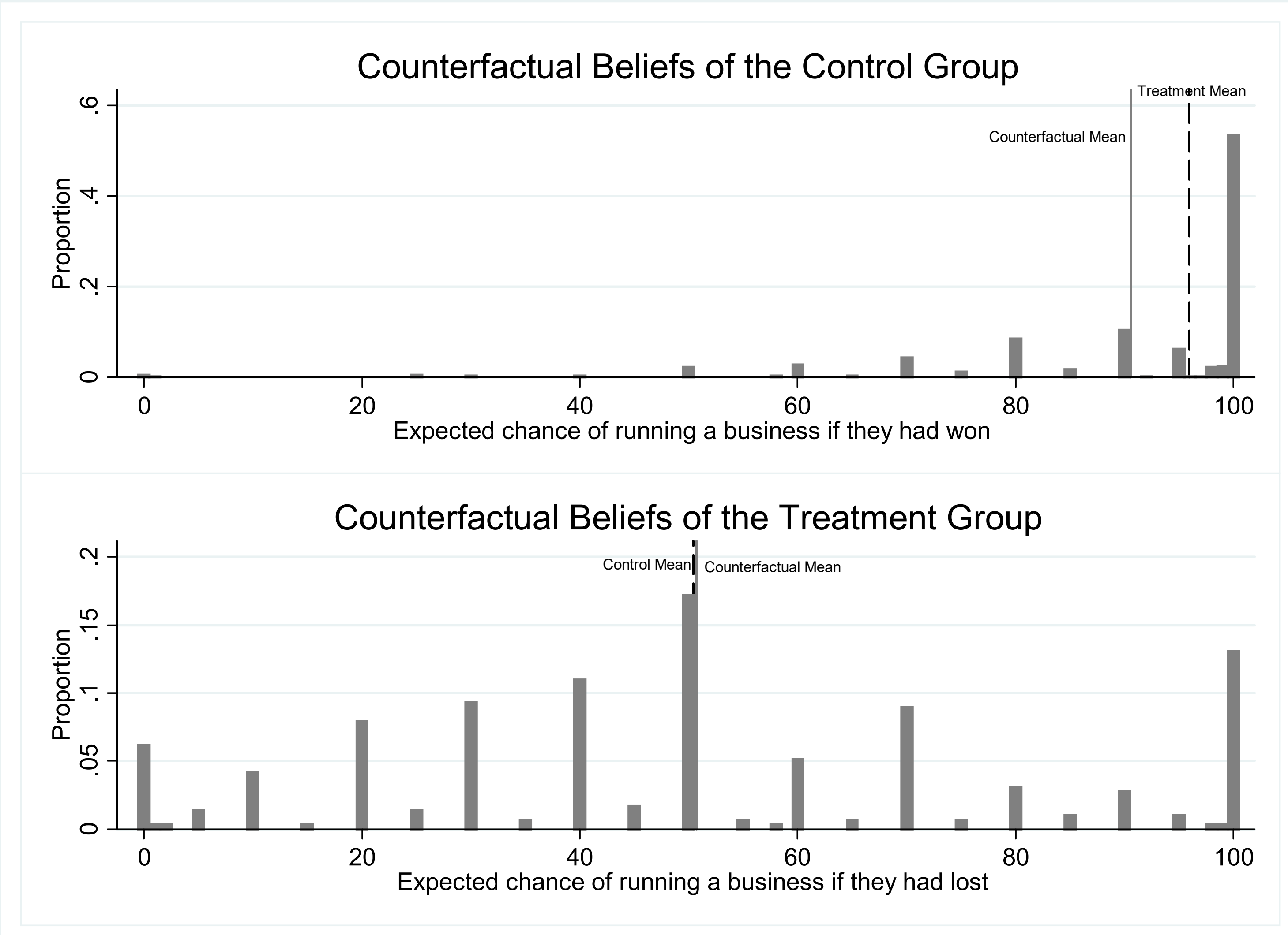


Figure 2: Counterfactual Beliefs of Likelihood of Operating a Firm for Existing Firm Applicants

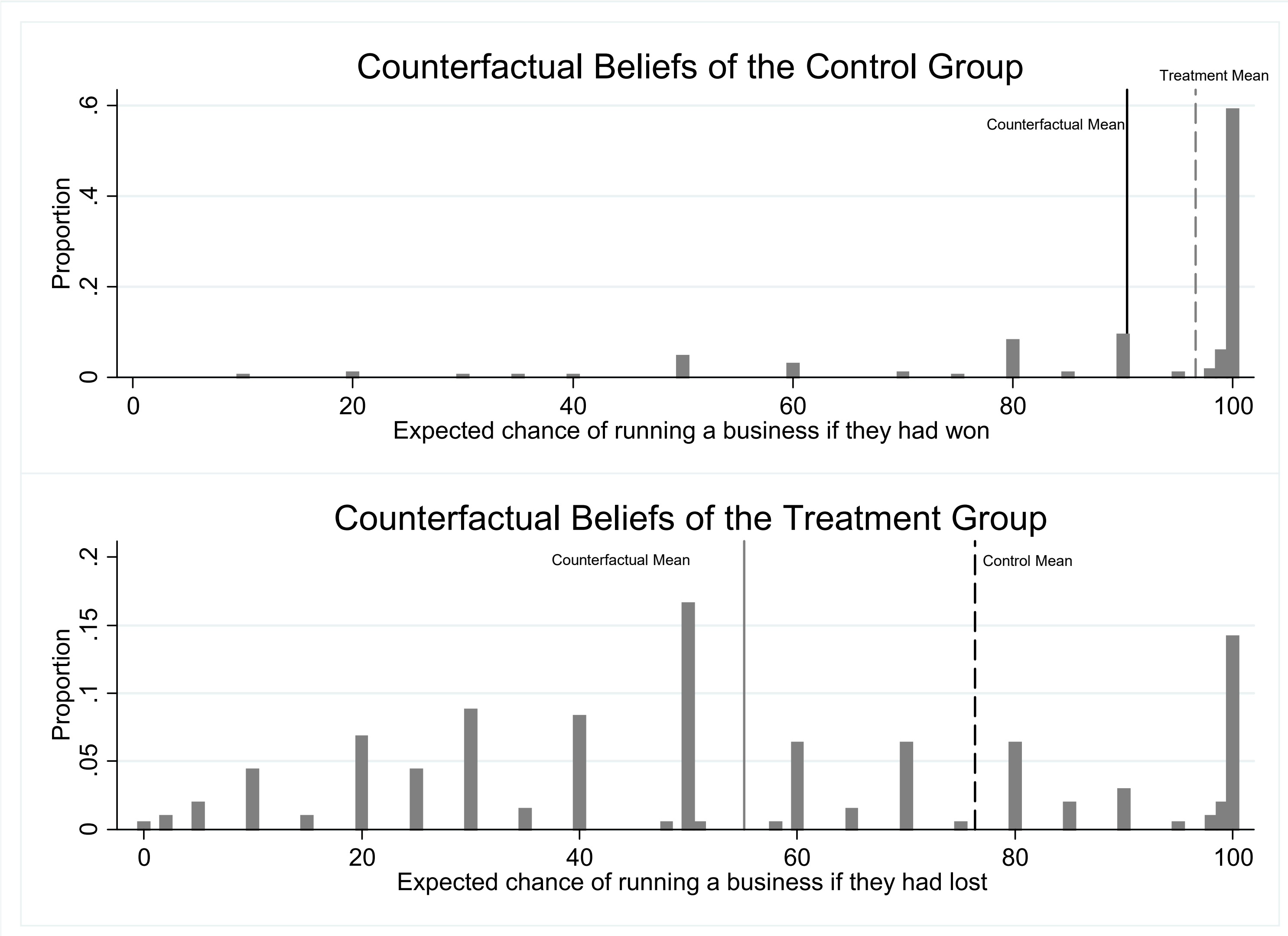


Figure 3: Comparison of Actual Distribution of Employment for Winners to Counterfactual Distribution expected by control group had they won

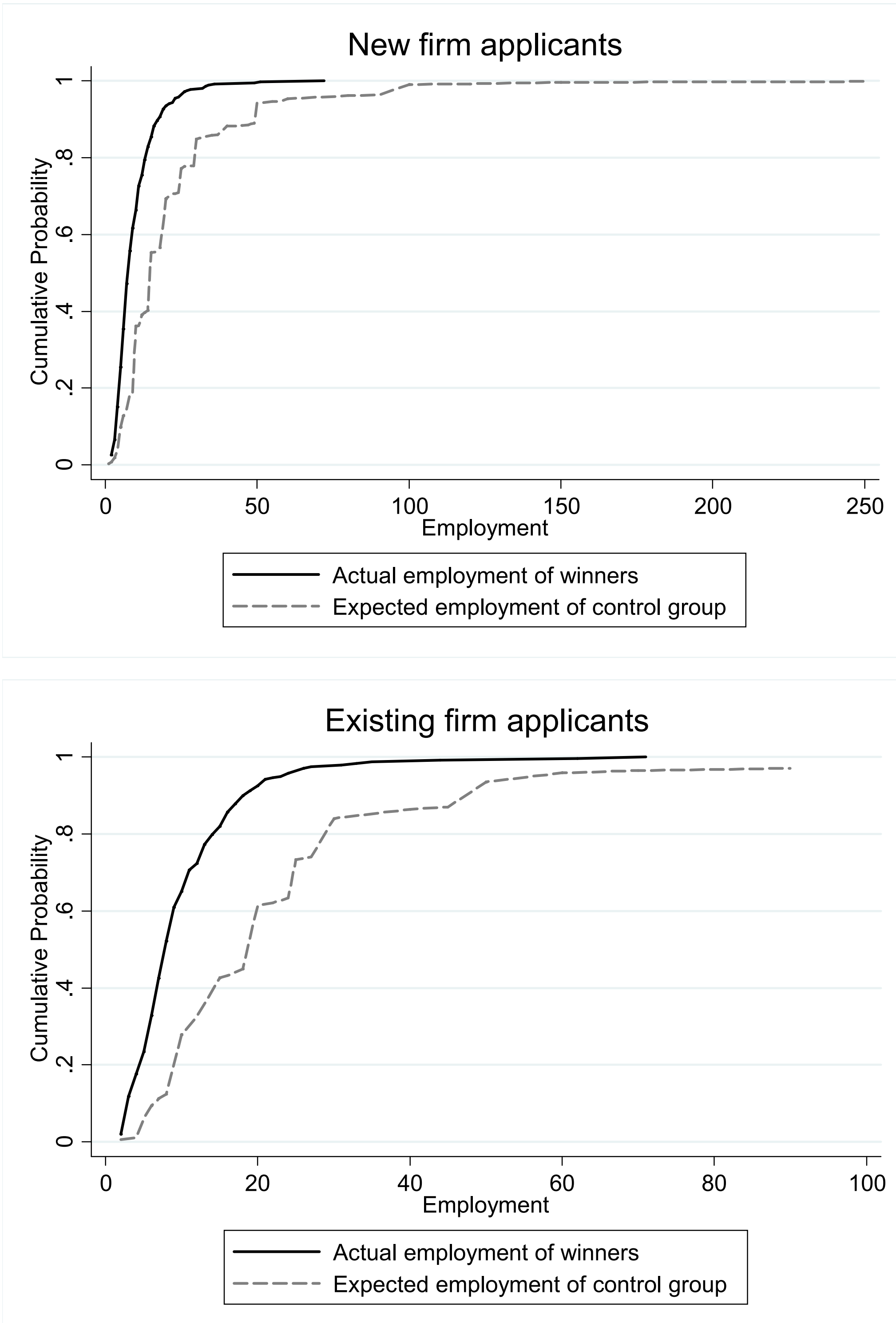
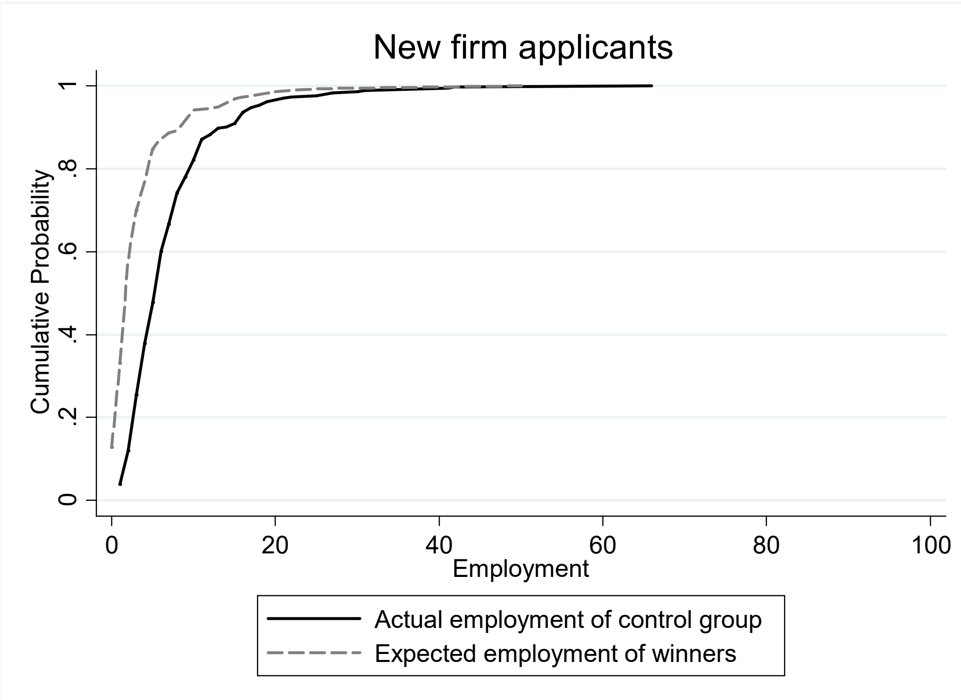


Figure 4: Comparison of Actual Distribution of Employment for Control Group to Counterfactual Distribution expected by treatment group had they lost



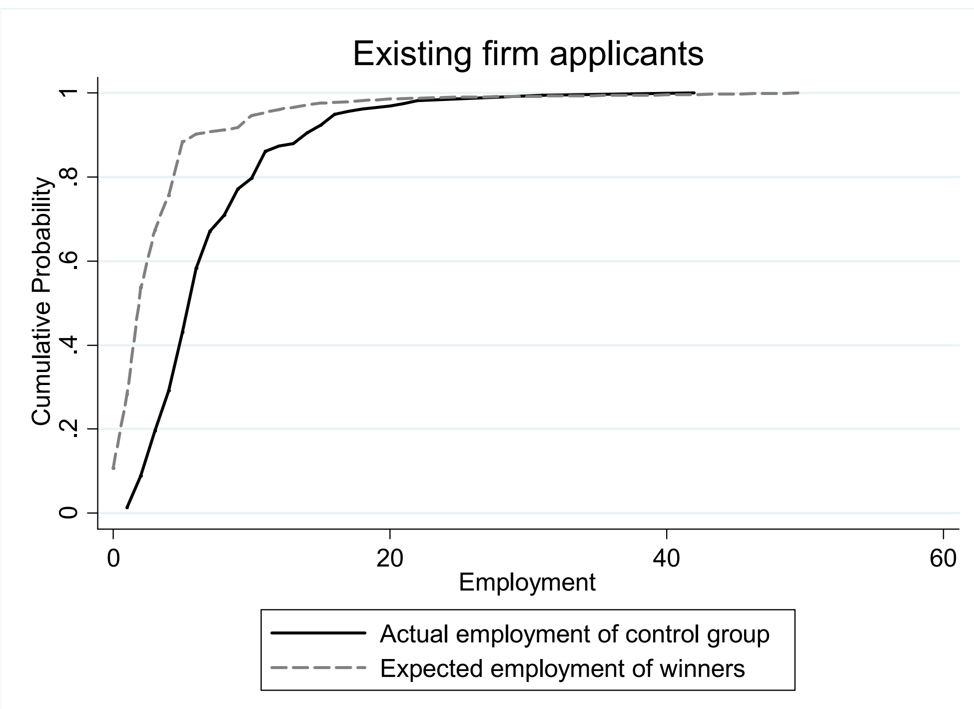
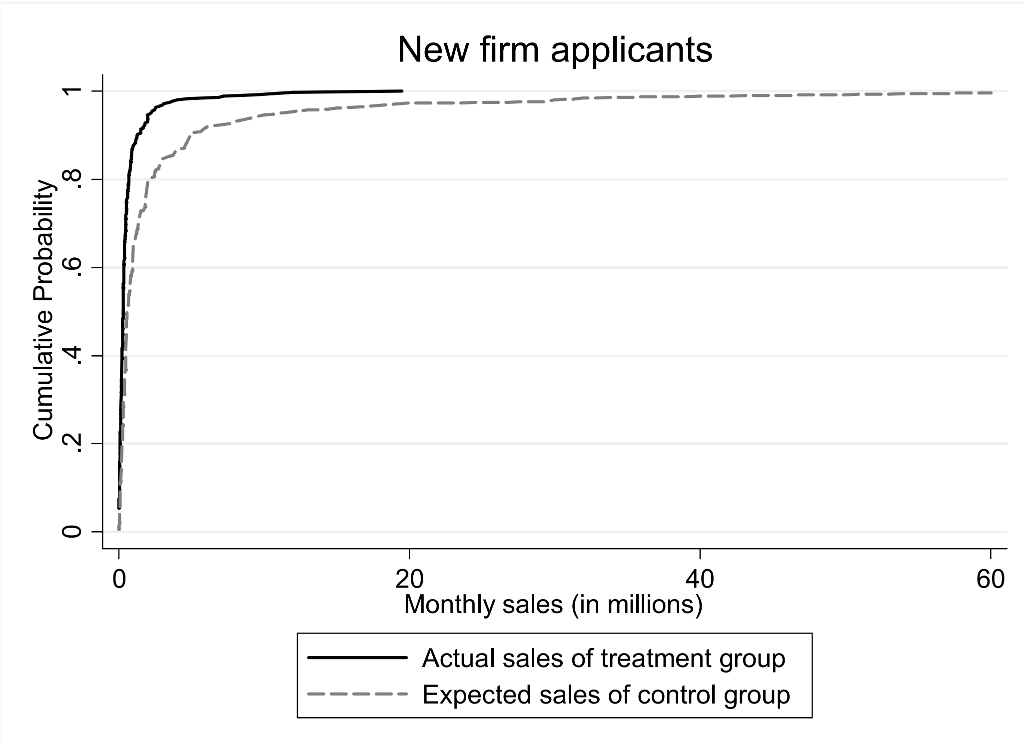


Figure 5: Comparison of Actual Distribution of Sales for Winners to Counterfactual Distribution expected by control group had they won



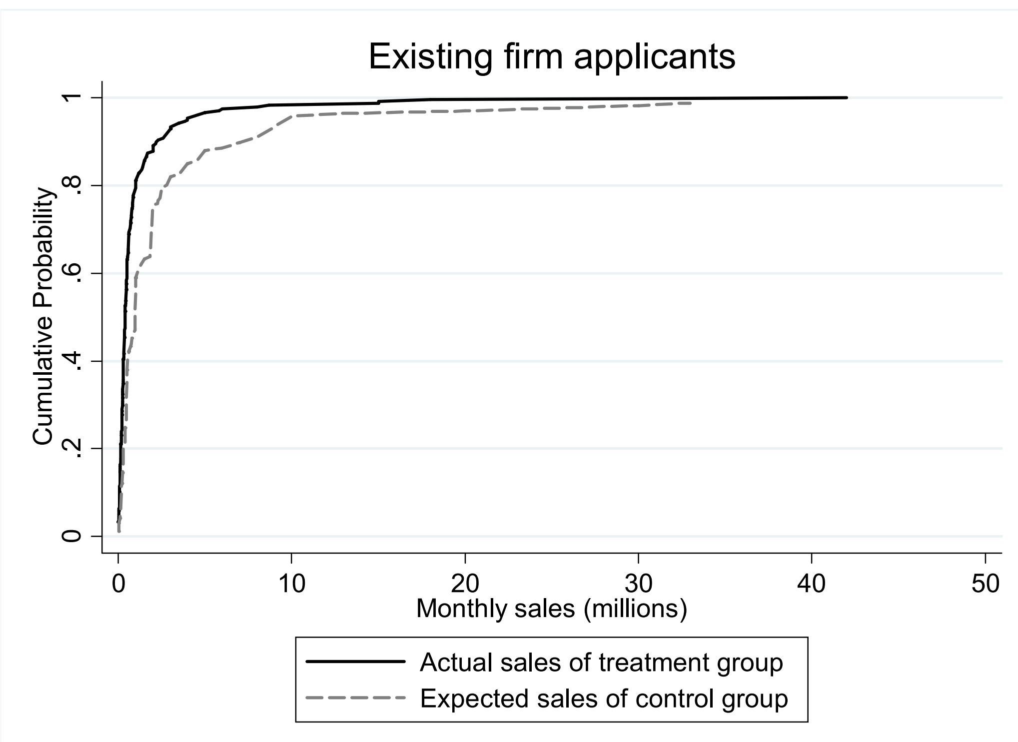


Figure 6: Comparison of Actual Distribution of Sales for Control Group to Counterfactual Distribution expected by treatment group had they lost

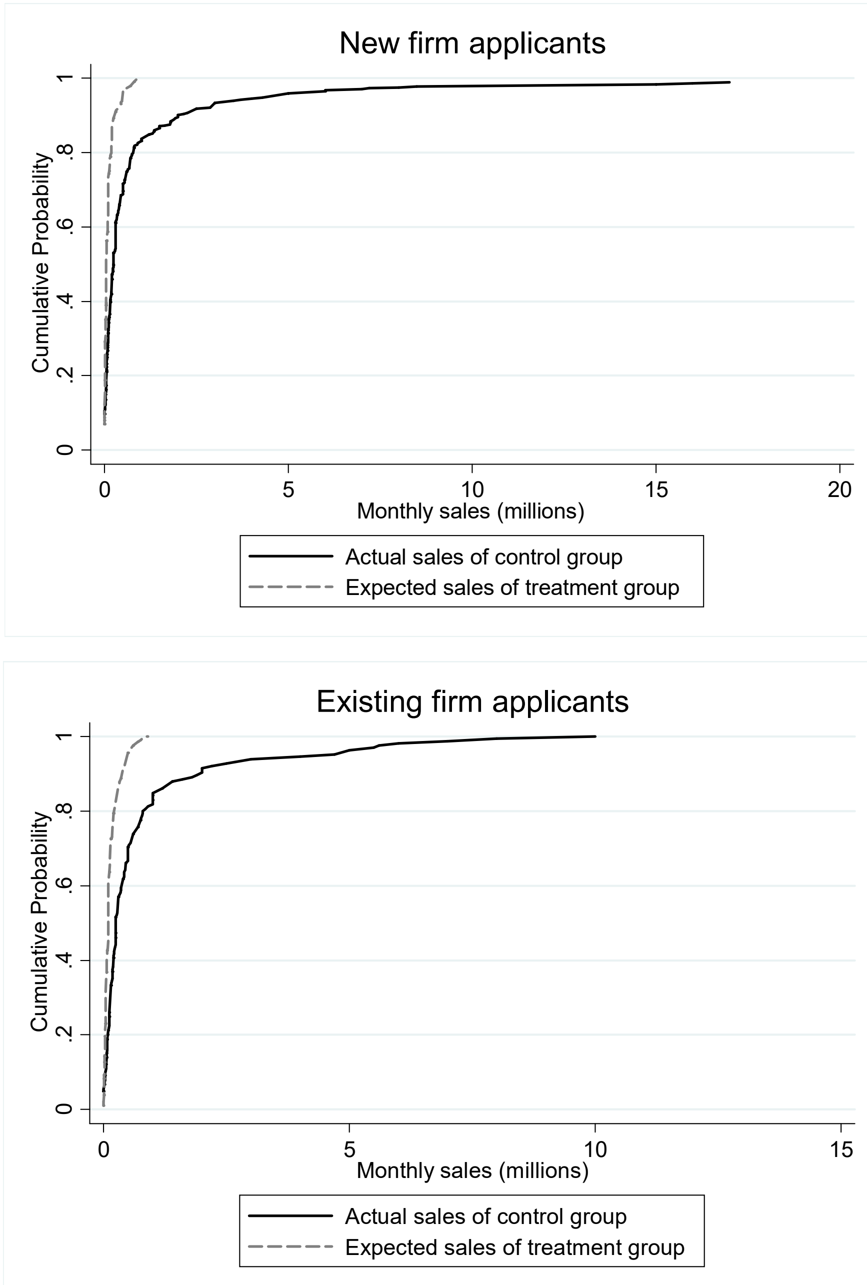


Table 1: Counterfactuals of the likelihood of operating a business

	Sample Size	Realized Outcome (Percent in Business)	Counterfactual Outcome (Percent chance of business if in alternative treatment) Mean (s.d.)	Median	Treatment Effect according to counterfactual	Experimental Treatment Effect and 95% interval.
Control Group Firms						
New Applicants	516	50.4	90.6 (16.4)	100	40.2	37.3 [32.6, 42.0]
Existing Applicants	169	76.3	90.4 (18.3)	100	14.1	19.6 [13.5, 25.7]
Treatment Group Firms						
New Applicants	291	95.9	50.7 (29.9)	50	45.2	37.3 [32.6, 42.0]
Existing Applicants	205	96.6	55.1 (29.5)	50	41.5	19.6 [13.5, 25.7]

Notes: Last column experimental estimates are from Table 2 of McKenzie (2015)

Table 2: Counterfactuals of the number of employees conditional on operating a business

	Sample Size	Realized Outcome (Mean number of employees)	Counterfactual Outcome Number of employees if in alternative treatment) Mean (s.d.)	Trimmed mean (s.d.)	Median	Treatment Effect according to counterfactual	Experimental Treatment Effect and 95% interval.
Control Group Firms							
New Applicants	515	6.9	23.2 (31.4)	18.3 (12.5)	15	16.3	2.2 [1.1, 3.3]
Existing Applicants	169	7.6	23.9 (19.7)	20.9 (13.3)	20	16.3	2.5 [1.0, 4.0]
Treatment Group Firms							
New Applicants	290	9.7	3.8 (5.6)	3.0 (3.1)	2	5.9	2.2 [1.1, 3.3]
Existing Applicants	205	10.2	3.7 (5.5)	2.9 (2.4)	2	6.5	2.5 [1.0, 4.0]

Notes: Last column experimental estimates are from appendix Table 11a of McKenzie (2015)
Trimmed mean drops counterfactuals above the 95th percentile

Table 3: Heterogeneity in Control Group New Applicant counterfactuals of the number of employees conditional on operating a business

	Sample Size	Realized Outcome (Mean number of employees)	Counterfactual Outcome Number of employees if in alternative treatment) Mean (s.d.)	Trimmed mean (s.d.)	Median	Treatment Effect according to counterfactual	Experimental Treatment Effect and 95% interval.
All Control Group New Applicants	515	6.9	23.2 (31.4)	18.3 (12.5)	15	16.3	2.2 [1.1, 3.3]
Females	86	5.8	17.8 (13.9)	16.9 (11.5)	15	12.0	3.3 [-0.1, 6.7]
Other Cities	408	7.1	22.6 (31.6)	18.0 (12.4)	15	15.5	2.7 [1.5, 3.9]
No-university	170	7.4	23.5 (27.7)	19.0 (13.5)	15	16.1	2.8 [0.9, 4.6]
University education	345	6.7	23.1 (33.1)	18.0 (12.0)	15	16.4	2.5 [1.1, 3.9]
Digitspan recall >=7	279	7.1	24.1 (35.6)	19.0 (13.3)	15	17.0	3.3 [1.7, 4.8]
Males	429	7.1	24.3 (33.8)	18.6 (12.7)	15	17.2	2.4 [1.3, 3.6]
Digitspan recall <=6	156	7.1	24.5 (29.4)	18.3 (12.4)	15	17.4	2.1 [0.5, 3.7]
Abuja and Lagos	107	6.5	25.4 (30.5)	19.5 (13.1)	15	18.9	1.2 [-1.3, 3.7]

Notes:Trimmed mean drops counterfactuals above the 95th percentile