

Unpacking the determinants of entrepreneurship development and economic empowerment for women*

PRE-ANALYSIS PLAN

February 2014

Principal Investigators

David McKenzie, *World Bank*

Silvia Paruzzolo, *ILO*

This plan and the associated trial were registered on the AEA Social Science Registry on February 21, 2014. The assigned registry number is AEARCTR-0000287

<http://www.socialscienceregistry.org/trials/287>

* We gratefully acknowledge funding provided to this project from 3ie and the ILO. Human Subjects Approval was obtained from Innovations for Poverty Action (13February-002) and the Maseno University Ethics Review Committee (MSU/DRPC/MUERC/000006/13). Authority to conduct research in Kenya was provided by the Kenyan Ministry of Science and Technology (NCST/RCD/14/013/553B).

1. Introduction

This impact evaluation aims to measure the causal impact of the ILO's *Get Ahead* business training programme on the profitability, growth and survival of female-owned businesses, and to evaluate whether any gains in profitability come at the expense of other business owners. To do so, the evaluation will use a randomized control trial (RCT) methodology with a two-level randomized experiment: randomized selection of villages, and of individuals within villages.

This pre-analysis plan outlines the hypotheses to be tested and specifications to be used in the analysis of the impact of the Get Ahead program. Since the authors completed the plan before the follow-up data was collected and analyzed, the plan can serve as a commitment and marker of the key outcomes to be examined, and help guard against the risks of data mining and multiple hypothesis testing.

The plan is outlined as follows: Section 2 reviews the motivation for the study, the sample selection and data sources; Section 3 outlines the theory of change; Section 4 enumerates the hypotheses to be tested as part of the study; Section 5 outlines the specifications to be used in analyzing the data. More details of the study are also contained in the baseline report.

2. Overview of the Study

2.1 Motivation and Program Description

Business training is one of the most common forms of active support provided by Governments, NGOs, and some international organizations, to small firms around the world. However, until recently there has been very little rigorous evidence as to the impacts of these programmes. A recent literature summarized in McKenzie and Woodruff (forthcoming) has begun to measure these impacts, but have often suffered from low statistical power, and have been unable to measure the extent to which gains to trained firms come at the expense of other firms.

The International Labour Organizations' Women's Entrepreneurship Development programme has introduced a business management training package with a strong gender component called *Gender and Enterprise Together (GET ahead)* in over a dozen countries. This program "differs from conventional business training materials by highlighting essential entrepreneurial skills from a gender perspective, whether applied to starting or improving an individual, family or group business. It addresses the practical and strategic needs of low-income women in enterprise by strengthening their basic business and people management skills. It shows women how to develop their personal entrepreneurial traits and obtain support through groups, networks and institutions dealing with enterprise development" (Bauer et al, 2004). The program began in Thailand in 2001, and over the next decade was used in 18 countries around the developing world.¹

An objective of the program is to create a "business mind" among low-income women engaged in small-scale businesses. The training methodology is participatory, with practical exercises to teach concepts. For example, women learn about the different types of costs involved in production, and how to account for their own costs through making lemonade; have role play exercises to practice different sales strategies for customers; and make necklaces to discuss a production process and the importance of different factors in product design.

2.2 Sample Selection and Assignment to Treatment

We summarize the process discussed in more detail in the baseline report. A Stakeholder retreat in October 2012 was used to pre-select 10 counties from the 47 counties in Kenya as possible locations for the study. A more detailed review of these 10 counties was then used to select 4 counties in which to provide training: Kakamega and Kisii in the Western region, and Embu and Kitui in the Eastern region.

In each of the four counties, between June and November 2013 field staff from Innovations for Poverty Action, Kenya, mapped out all market centers deemed as medium or large outside of the main cities. Field staff then conducted a market census, applying a 31 question, 7 page listing questionnaire to each female-owned enterprise operating on a non-market day in these markets.

¹ http://www.ilo.org/wcmsp5/groups/public/---ed_emp/documents/publication/wcms_159071.pdf

We then applied an eligibility filter to determine which women to include in the baseline survey. This filter required the women to have reported profits, and not to have reported profits that exceeded sales; to have a phone number that could be used to invite them for training; to be 55 years or younger in age; to not be running a business that only dealt with phone cards or m-pesa, or that was a school; that the person responding not be an employee; that the business not have more than 3 employees; that the business have profits in the past week between 0 and 4000 KSH; that sales in the past week be less than or equal to 50,000 KSH; and that the individual had at least one year of schooling. These criteria were chosen to reduce the amount of heterogeneity in the sample (thereby increasing our ability to detect treatment effects), and to increase the odds of being able to contact and find individuals again.

Baseline surveys took place soon after the listing surveys in each county, between June and November 2013. Out of a target of 4,037 individuals, we were able to interview 3,538 (87.6%) in time to consider them for inviting to training.

After dropping one more market with too few individuals, we were left with a sample of 3,538 individuals in 157 markets who had satisfied the screening criteria and completed the baseline survey. These individuals were then assigned to treatment and control in a two-stage process.

First, **Markets** were assigned to treatment (have some individuals in them invited to training) or control (no one in the market would be invited to training) status. Randomization was done within 35 strata defined by geographical region (within county) and the number of women surveyed in the market. In Kakamega and Kisii the size strata were 26 or under, 27 to 30, 31+; in Embu the size strata were <20 or 20+, and in Kitui the size strata varied within smaller geographic clusters between the two cuts used elsewhere. The need to ensure sufficient numbers for training meant that more than half the markets were allocated to treatment, with 93 markets assigned to treatment and 64 to control (see table below).

Then within each market, **Individuals** were assigned to treatment (be invited to training) or control (not be invited to training) within treated markets by forming four strata, based on quartiles of weekly profits from the census (≤ 450 , 451-800, 801-1500, 1501-4000), and then assigning half the individuals within each strata to training. When the number of individuals in the strata was odd, the odd unit was also randomly assigned to training. This resulted in 1173 of

the 2161 individuals in treated markets being assigned to treatment, and 988 to control (see table below).

Summary of Random Assignment

County	# of Strata	Markets		Individuals		
		Treatment	Control	Treatment in Treated Market	Control in Treated Market	Control in Control Market
Kakamega	9	19	12	257	220	305
Kisii	9	20	14	275	237	333
Embu	6	20	16	231	189	295
Kitui	11	34	22	410	342	444
Total	35	93	64	1173	988	1377

2.3 Training dates and take-up

Training took place during the following dates:

- Kakamega: June 25-July 20, 2013
- Kisii: July 19-September 6, 2013
- Embu: September 16-October 11, 2013
- Kitui: October 28-November 29, 2013

Of the 1173 individuals assigned to training, 87.3% said they would attend when invited, and 911 (77.7%) attended at least one day of training. Of the individuals who attended at least one day, 94.6% attended all 5 days.

Key data sources

The primary sources of data are:

- 1) Data from the **market census** conducted pre-baseline. This contains basic information on the business, including the key variables used to form randomization strata.

- 2) Data from the **baseline survey** conducted before random assignment and before training.
- 3) Data from **a first follow-up survey (currently scheduled to begin June 2014, corresponding to one year since the first training began);** and from a **second follow-up survey (currently scheduled to begin June 2015).**

This will be supplemented by i) administrative data on training take-up; and ii) data on the observed value of inventories and business assets based on independent valuations of photos taken of the businesses,

3. Theory of Change

The theory of change to be tested is as follows:

Step 1: When training is available, individuals take it up. This will be measured through take-up rates.

Step 2: Individuals stay in the training, and improve their business knowledge as a result of the training. This will be measured through questions on business knowledge embedded in the follow-up surveys.

Step 3: As a result of this additional business knowledge, women will implement some of the ideas and practices that the training teaches. This will be measured through the follow-up surveys and treatment effects measured.

Step 4: Implementation of these practices will improve business profits and revenues. The key question left open by existing work is how this step occurs. Various explanations have been suggested, including

- (a) Increased marketing and better presentation of the business leads customers to move from non-treated to treated firms.

- (b) Better record-keeping enables businesses to keep better track of business conditions and react more quickly to down periods (Drexler et al, 2011; Karlan and Valdivia, 2011), so that the difference comes about through performance in bad periods.
- (c) Better record-keeping and business management enables firms to be in a position to plan their finances better and obtain more financing from banks and microfinance organizations, resulting in higher production as a result of more capital.
- (d) Better inventory management and better purchasing practices enable firms to reduce spoilage of stock and taking advantage of bulk discounts, thereby lowering business expenses.
- (e) Business training directed at women teaches them strategies to avoid household and family demands that otherwise constrain their ability to reinvest profits in their business.
- (f) Trained women have greater self-efficacy in entrepreneurial tasks, and devote more time to working in their business.

Step 5: Improved business outcomes lead to real improvements in women’s lives. Some studies have suggested that economically advancing women may have positive impacts on their empowerment in their homes and communities, but that it could also have some unintended negative consequences such as leading to increased conflict in the household, increased time pressure, or loss of business control. The second round of qualitative interviews will address these questions.

4. Hypotheses to be Tested

We will test this theory of change through a set of hypotheses, which are presented in groups, summarized in the table below.

Hypothesis Groups

A.	Impact on Primary Outcomes for Treated Firms: GET Ahead training may have positive average impacts on profitability, sales, and business survivorship for treated firms.
B.	Impact on Secondary Outcomes for Treated Firms: GET Ahead training may have positive effects on overall employment, empowerment, subjective well-being, and household asset ownership.
C.	Impact on Primary Outcomes for Non-treated Firms: Having other firms in your marketplace receive training may have negative impacts on the profitability, sales, and survivorship of non-treated firms.
D.	Causal Chain of Process and Mechanisms: Training should lead to increases in business knowledge, more deployment of business practices, and potentially operate to increase profitability or sales through increased marketing, better responsiveness to down periods, greater access to finance, better inventory management, better avoidance of household demands, increased self-confidence and attitudes, and higher social capital.
E.	Heterogeneity of Impacts: Training should be more effective for higher skilled women and more empowered women, and should have greater spillovers for firms facing more competition

GROUP A: Impact on Primary Outcomes for Treated Firms

Hypothesis A1: Treatment leads to higher survivorship.

This will be measured by the following outcome:

1. **Business is still operating at the time of the survey.** This will be measured directly for those answering the follow-up survey, and assessed via interviewer observation and third-party reporting for those unable to be re-interviewed. Since all firms were in existence at baseline, the baseline value of the dependent variable will not be included here.

One reason this hypothesis may not hold is that training may also lead less skilled and less profitable business owners to realize they may be better off closing their businesses. However, conversely these may be the very types of firms that training helps survive when they otherwise wouldn't. We will examine this by testing the heterogeneity of treatment effect with respect to initial profitability, and initial business skills. This leads to:

Hypothesis A1-H: The treatment effect on survivorship will vary with:

- a) Initial profits (measured by whether profits in the initial market census were above 1000 KSH per week or not; 46% of firms had profits above this)
- b) Baseline business practice score (measured by whether the baseline score is above the median of 11 practices out of a maximum 28 point score).

Hypothesis A2: Treatment leads to business owners making higher sales, and earning greater profits.

This will be measured as the following set of outcomes:

1. **Total sales in the last day, truncated at the 99th percentile.** (Baseline Q5.6a).
2. **Total sales in the last week, truncated at the 99th percentile.** (Baseline Q5.6b)
3. **Total sales in the last week of product accounting for largest share of profits, truncated at the 99th percentile.** (Baseline Q5.8d (price per unit)*Baseline Q5.8e (number of units sold)).
4. **Average weekly profits over the last 2 weeks, truncated at the 99th percentile** (Measured at market census). If they only report profits for one of the last two weeks and the other is missing, the week for which data is recorded will be used.
5. **Mark-up profits in the last week of product accounting for largest share of profits, truncated at the 99th percentile.** (Baseline Q5.8d (price per unit) – 5.8c (cost per unit))*(Baseline Q5.8e (number of units sold)).
6. **Third-party valuation of business inventories for sale, truncated at the 99th percentile** (baseline administrative data from photos)
7. A standardized **profits and sales** impact will be obtained by aggregating these different effects as described below in our methods section as a standardized z-score.

Measurement of sales and profits in small firms is difficult, and there are several issues that we will need to deal with. Our planned approach to deal with several key issues is as follows:

1. Large heterogeneity across firms: the initial sample was screened and selected to be relatively homogenous in terms of weekly profits. Given this, and the likely non-trivial number of firms earning zero sales or profits in a particular month or week, we will use levels rather than logs. Nevertheless, as a robustness check, we will also examine the impact of treatment on the inverse hyperbolic sine transformation of total profits in the last month, and total sales in the last week: $\log(y+(y^2+1)^{1/2})$ – which is similar to the log transformation, but can deal with zeros.
2. Item non-response: the baseline survey had disappointingly high item non-response rates for profits, along with some item non-response for sales. This will be addressed in several ways. First, better training of enumerators will stress techniques for better eliciting such information. Second, the use of netbooks and survey software, rather than paper surveys, in the follow-up round can help prompt better for this information. Third, for those individuals refusing to give point estimates, range questions will be used. For businesses not answering the exact answer, but answer the range question, the midpoint of the range will be used. For firms in the top range, a value equal to the median of firms with sales in this top range will be used. Missing baseline values will be then dummied out in ANCOVA specifications. Finally, our field team suggested that Kenyan business owners will find it easier to report on the last week rather than last month, so we have moved to asking about weekly profits.
3. Dealing with firms which close down: Since treatment may affect both the proportion of firms which shut down, as well as the selectivity of which firms close down, simply comparing profits or sales for surviving firms may be biased. Our primary approach will therefore be to record profits and sales as zero for firms which are closed. If we find no significant impacts of treatment on either the rate or selectivity of survivorship, we will also conduct exploratory analysis which looks at impacts of treatment conditional on surviving.

Hypothesis A3: Treatment does not affect reporting errors

There are two concerns here. The first is that with any program involving business training or improvements in record-keeping, it may lead to changes in the accuracy of the information being reported, even if the underlying business financial position does not change. If businesses

systematically under- or over-state sales and profits, this will lead to a bias in the measured treatment effect. A second concern is that participants may wish to show gratitude for being given training, and exhibit desirability bias, giving biased responses they think that interviewers wish to hear.

We will address this issue in a couple of ways. The first is to test whether the training has affected the reporting of firms through estimating the treatment impact on the number of reporting errors made. The following will be deemed a reporting error, and our measure will be the total number of such errors:

- a) Revenues in last week from main product exceed total revenues in the last week ($Q5.8d * Q5.8e > Q5.6b$)
- b) Profits in the last week from main product exceed total profits in the last week (not measured at baseline)
- c) Total profits in worst week of 2014 are higher than total profits in the last week (not measured at baseline)
- d) Profits in the last week exceed sales in the last week (not measured at baseline)
- e) Total sales yesterday exceed total sales in the last week ($Q5.6a > Q5.6b$)
- f) Cost of the raw materials used to produce one unit costs more than the price charged per unit for the most profitable good ($Q5.8d < Q5.8c$)

Second, we will examine whether treatment significantly affects the correlation between an independent measure of inventory levels (taken from valuing goods for sale in a photo) and the reported valuation of inventory levels.

Third, as a robustness check, we will examine how sensitive the impact of treatment on truncated total profits in the last month, and truncated total sales in the last week, is to controlling for the record-keeping score from our business practices ($\text{recordscore} = \text{rsum}(b_s9_14 - b_s9_20 \ b_s9_22)$ in baseline).

GROUP B: Impact on Secondary Outcomes for Treated Firms

GET Ahead training may have positive effects on overall employment, empowerment, subjective well-being, and household asset ownership.

Hypothesis B1: Treatment leads to individuals being more likely to be employed, and earning higher income from labor.

This will be measured as the following set of outcomes:

1. **Employed for pay:** this will be coded as 1 if they still work as a business owner, or if they now work as a wage worker. For individuals unable to be re-interviewed, this will be obtained via proxy reporting.
2. **Income from work:** this will be the sum of any profits earned in the business in the last week, plus any income from wage work.

Hypothesis B2: Treatment leads to higher empowerment of women in terms of decision-making around finances and business:

This will be measured as the **sum** of the following set of binary outcomes:

1. **Compelled to spend money on husband or family** (Baseline 11.1) – coded as 1 if they answer no, zero otherwise. 8.7 percent answer no at baseline.
2. **Not the only person with access to business money** (Baseline 11.2) – coded as 1 if only myself has access to the money made in the business, 0 otherwise. 90.1 percent answer yes at baseline.
3. **Has some money which they have sole control over and can spend how they like** (Baseline 11.4), coded 1 if yes, 0 otherwise: 89.6 percent answer yes at baseline.
4. **Do not need anyone's permission to visit a friend.** (Baseline 11.7_1 = 1 or 2), 79% do not at baseline.
5. **Do not need anyone's permission to travel to sell a business asset** (Baseline 11.7_2 =1 or 2), 86% do not at baseline
6. **Do not need anyone's permission to travel to a new location to work** (Baseline 11.7_3=1 or 2), 68% do not at baseline
7. **Do not need anyone's permission to stay overnight in a different town** (Baseline 11.7_4=1 or 2), 50% do not at baseline

8. **Do not need anyone's permission to work later than usual hours** (Baseline 11.7_5=1 or 2), 81% do not at baseline.
9. **Do not need anyone's permission to take out a loan** (Baseline 11.7_6=1 or 2), 57% do not at baseline.
10. **Do not need anyone's permission to spend money on an investment for their business** (Baseline 11.7_9=1 or 2), 90% do not at baseline.

Hypothesis B3: Treatment leads to an increase in subjective well-being and improvement in mental health.

This will be measured as the following set of outcomes:

1. Subjective standard of living as measured by a standard Cantril ladder of life question (not measured at baseline)
2. Anticipated Subjective standard of living in 5 years on Cantril ladder (not measured at baseline)
3. Mental health score, as measured by the MHI-5 index (not measured at baseline)

Hypothesis B4: Treatment will lead to an increase in household durable asset ownership.

This will be measured as the number of a set of 10 household durable assets owned by the household (Baseline 10.3).

GROUP C: Impact on Primary Outcomes for Non-Treated Firms

Hypothesis C1: Having other firms in your marketplace receive training may have negative impacts on the profitability, sales, and survivorship of non-treated firms.

This will be investigated using the same outcomes used for treated firms:

1. **Business is still operating at the time of the survey.** This will be measured directly for those answering the follow-up survey, and assessed via interviewer observation and third-party reporting for those unable to be re-interviewed. Since all firms were in existence at baseline, the baseline value of the dependent variable will not be included here.

2. **Total sales in the last day, truncated at the 99th percentile.** (Baseline Q5.6a).
3. **Total sales in the last week, truncated at the 99th percentile.** (Baseline Q5.6b)
4. **Total sales in the last week of product accounting for largest share of profits, truncated at the 99th percentile.** (Baseline Q5.8d (price per unit)*Baseline Q5.8e (number of units sold)).
5. **Average weekly profits over the past two weeks, truncated at the 99th percentile** (market census provides baseline value)
6. **Mark-up profits in the last week of product accounting for largest share of profits, truncated at the 99th percentile.** (Baseline Q5.8d (price per unit) – 5.8c (cost per unit))*(Baseline Q5.8e (number of units sold)).
7. **Third-party valuation of business inventories for sale, truncated at the 99th percentile** (baseline administrative data from photos)
8. A standardized **profits and sales** impact will be obtained by aggregating the different effects (2-7) as described below in our methods section as a standardized z-score.

Hypothesis C2: The benefits to treated firms outweigh the costs to non-treated firms in the same villages, so that total sales and profits of sampled firms in treated market places increase.

This will be estimated via a marketplace level regression, using the following outcomes:

1. **Average sales in the last week in the market among sampled firms.** Constructed as the average of sales in the last week (truncated at the 99th percentile) among all firms in our sample from that market.
2. **Average profits in the last two weeks in the market among sampled firms.** Constructed as the average of profits in the last two weeks (truncated at the 99th percentile) among all firms in our sample from that market.

Firms which are closed down will be coded as having zero profits and zero sales for the purpose of these calculations. Firms with item non-response for profits or sales will be excluded from forming the village average.

GROUP D: Causal Chain of Process and Mechanisms

Training should lead to increases in business knowledge, more deployment of business practices, and potentially operate to increase profitability or sales through increased marketing, better responsiveness to down periods, greater access to finance, better inventory management, better avoidance of household demands, and increased self-confidence.

Hypothesis D1: Treatment will lead to an increase in Business Knowledge, and an increase in the use of standard Business Practices

This will be measured via the following outcomes:

1. **Number of Business Knowledge Questions Correct:** out of 7 questions to be asked on follow-up survey (not asked at baseline)
2. An **index of business practices** formed from section 9 at baseline, aggregating the number of practices firms carry out, using the same coding as used by de Mel et al. with these questions in Sri Lanka (see appendix of de Mel et al. business training paper).

Hypothesis D2: Increased marketing and better presentation of the business leads treated firms to gain more customers, potentially from non-treated firms.

This will be measured via the following outcomes:

1. **Number of customers in a typical week – impact on treated firms** (not measured at baseline). This will be top-coded at the 99th percentile of the overall distribution to account for outliers.
2. **Number of customers in a typical week – impact on non-treated firms** (not measured at baseline). This will be top-coded at the 99th percentile of the overall distribution to account for outliers.
3. **Have gained a new regular customer in the past three months who previously did not shop at your business – impact on treated firms** (not measured at baseline)
4. **Have lost a regular customer in the past three months – impact on non-treated firms** (not measured at baseline)

Hypothesis D3: Better record-keeping enables businesses to keep better track of business conditions and react more quickly to down periods

This will be measured via the following outcomes:

1. **Ratio of Business Profits in Worst Week of 2014 to Business Profits in Current Week**
2. **Business regularly uses business records to know how sales of a particular product are increasing or decreasing from one month to another.** Baseline Q9.17=1.

Hypothesis D4: Treated firms are able to obtain more financing from banks and microfinance organizations, and build up higher levels of inventories and capital stock.

This will be measured via the following outcomes:

1. **Firm has obtained a loan in the last year from a private bank, government bank, or microfinance organization** (control for baseline ever have had a loan from these entities, Q4.4_1-Q4.4_3)
2. **Value of Inventories and Raw Materials, truncated at the 99th percentile** (baseline Q5.3)
3. **Total value of business tools, machinery, equipment, furniture, and vehicles (capital stock excluding land and buildings), truncated at 99th percentile** (baseline Q5.1_1-Q5.1_4).

Hypothesis D5: Better inventory management and better purchasing practices enable firms to reduce spoilage of stock and taking advantage of bulk discounts, thereby lowering business expenses.

This will be measured via the following outcomes:

1. **Fraction of Business Inventory Lost to Spoilage in the Last Month** (not measured at baseline)
2. **Did the business receive a bulk discount when purchasing goods in the past month?** (not measured at baseline)
3. **Percentage change in the cost of producing one unit of most profitable product compared to baseline** (baseline Q5.8c). This will be truncated at the 5th and 95th percentiles, since we expect this to be quite noisy.

Hypothesis D6: Trained women have greater self-efficacy in entrepreneurial tasks, devote more time to working in their business, and have stronger entrepreneurial attitudes.

This will be measured via the following outcomes:

1. **Owner's labor hours:** measured as hours in the last week (baseline Q1.5a). This will be top-coded at the 99th percentile of the overall distribution.
2. **Entrepreneurial self-efficacy** – Measured as the number of business activities that the owner rates themselves as “very confident” in their ability to do (not asked at baseline). This is coded as 1 for each item if the owner answers 4 = very confident, and 0 if they answer 1 through 3, or 9 (not applicable or refuse).
3. **Get Ahead Attitudes:** Measured as the sum of scores on 11 questions (baseline 3.9a-3.9k) designed to measure attitudes GET Ahead is meant to encourage. These are scored 1 through 5, where 1=strongly disagree, 5 = strongly agree. Questions will be coded so that higher scores indicates better entrepreneurial attitudes. As a result, questions C, D, F and J will be reverse-coded. Individuals missing responses for particular questions will have their score out of the number of questions answered scaled up to be out of 11 questions.

Hypothesis D7: Trained women have greater social capital for business purposes, in the sense of stronger connections to other female entrepreneurs, more group negotiating, and being more likely to belong to a women's association.

This will be measured via the following outcomes:

1. **Belongs to a Women's Association:** This was measured at baseline, but it was revealed that women did not have a clear understanding of what was meant by a women's association, so we will not use baseline data here.
2. **Number of other women they discuss business matters with:** (follow-up 6.3). This will be truncated at the 99th percentile.
3. **Work together with other women when negotiating discounts or buying raw materials** (follow-up 6.4)

GROUP E: Key dimensions of Treatment Effect Heterogeneity

We are interested in three main dimensions of treatment effect heterogeneity, regarding how the effect of treatment on spillovers varies with the scope for competition, and how the effectiveness on individual firm owners varies with their existing skill levels and empowerment.

Hypothesis E1: The negative spillover effect of training on non-trained firms will be stronger for firms facing more competition, and for firms producing a higher share of their goods for the local market.

Testing this will entail testing treatment heterogeneity of sales in the last week, and profits in the last week, for the sample of control firms in treated and control markets, with respect to the following variables:

1. **Fewer than 5 other firms operate in the same line of business as them in their marketplace.** (Baseline Q8.1). 23.6 percent say this is true at baseline; assuming those who say they don't know how many firms are in their line of business are facing more than this number.
2. **At least 1 of 4 named individuals in the marketplace is classified as a competitor to their business** (Baseline Q8.6): Median is 1, 49.0% say they none of the 4 named people are competitors.
3. **Proportion of other firms surveyed in the same market that are in the same industry in our sample.** (Baseline s1_2_2_2). This is based on our codings of industry, and is a continuous variable.

Hypothesis E2: Training is more effective for women who initially have higher skills.

This will be tested by looking at the interaction of the training effect on profits in the last week for the sample of treated individuals and control individuals in control markets with respect to the following:

1. **Above median index of business practices score at baseline.** This index of business practices is defined above and in de Mel et al. (2014) and has a median of 11 points.
2. **12 or more years of education according to baseline data.** 28 percent have this.
3. **Digitspan recall above 5 digits in baseline.** 27 percent are above.

This is a priori unclear, since it depends whether training is a complement or substitute for existing skill levels. Our hypothesis is that initially more skilled women have more scope for growth, and that training may help them take advantage of this.

Hypothesis E3: Training is more effective for women with higher initial levels of empowerment.

This will be tested by looking at the interaction of the training effect on profits in the last week for the sample of treated individuals and control individuals in control markets with respect to the following:

1. **Above median on baseline empowerment measure**, defined as the sum of the 10 measures listed under hypothesis B2.

5. Estimation methodology

5.1 Individual Firm Level Analysis

The majority of our estimation will be at the individual firm level, and involve the following ANCOVA specification for firm i in market j :

$$Y_{i,j,t=1} = \beta_0 + \beta_1 T_{i,j} + \beta_2 S_{i,j} + \pi Y_{i,j,t=0} + \gamma M_{i,j,t=0} + X'_{k,i,j} \theta + \varepsilon_{i,j,t=1} \quad (M1)$$

Where $Y_{i,j,t=1}$ is the given outcome variable measured post-treatment, $Y_{i,j,t=0}$ is its baseline value and $M_{i,j,t=0}$ a dummy variable indicating whether or not this baseline value is missing, $T_{i,j}$ is an indicator for being in a treatment market and being assigned to treatment, $S_{i,j}$ is the spillover term, measuring whether firm i is a control firm in a market assigned to treatment; X_k is a vector of randomization strata dummy variables (geographic region*market size*profit range in market census) and $\varepsilon_{i,j,t}$ is the error term. β_1 will provide the intent-to-treat effect, which is the effect of being assigned to treatment relative to being a firm in the control markets, while β_2 measures the spillover effect by comparing control firms in treated markets to control firms in control markets. The standard errors will be clustered at the market level to account for the market level random assignment.

If there are no spillovers, we can also ask what is the impact of being assigned to training relative to not being selected for training within the same village? This will use the sample of treated firms in the treatment markets, and control firms in treatment markets, and estimate (M1) without the spillover term. Since training was randomized at the individual level within villages, this does not require clustering of standard errors, and Huber-White standard errors will be used. This specification offers greater power, but requires the assumption (tested above) of no spillovers.

Since not all those who were selected to be invited to training actually attended, we can also estimate the treatment effect on the treated by estimating:

$$Y_{i,j,t=1} = \beta_0 + \beta_1 C_{i,j} + \beta_2 S_{i,j} + \pi Y_{i,j,t=0} + \gamma M_{i,j,t=0} + X'_{k,i,j} \theta + \varepsilon_{i,j,t=1} \quad (M1)$$

where C_i is an indicator for attending the training course, which is instrumented by assignment to treatment status, T_i .

In cases where an outcome variable was not collected at baseline, these same specifications will be estimated without the control for baseline outcome.

5.2 Market Level Analysis

We will also estimate the overall impact of treatment assignment on sales and profits in the village market (hypothesis C2) by estimating the following equation at the level of the 157 markets:

$$\text{Average profits in market } j = a + b * \text{Market } j \text{ assigned to Training} + d' X_j + e_j$$

Where X_j are dummies for the market randomization strata (geographic region*market size).

5.3 Procedure for Utilizing Multiple Measurement

The specifications above are intended for use with the first follow-up survey. When multiple measures of follow-up outcomes are available after additional surveys are taken, we will test to see whether the treatment effects vary with time. If we cannot reject equality of treatment effects

over time, the multiple measurements will be combined to increase power following the procedures set out in McKenzie (2012).

5.4 Estimation of Heterogeneous Treatment Effects

Heterogeneous treatment effects will be estimated by interacting treatment status and the lagged dependent variable in (M1) or (M2) with the variable of interest Z.

5.5 Procedures for Dealing with Testing for Multiple Outcomes

We have a relatively rich set of outcome measures. To deal with multiple hypothesis testing we employ several approaches.

- 1) We specify a set of primary outcomes (Hypothesis A1 and A2, C1), which we see as the main outcomes of the study, and distinguish these from impacts on other outcomes (groups B and D of hypotheses).
- 2) We will look at standardized treatment effects within domains where it makes sense to combine measures into a single aggregate. To do this we follow the approach of Kling, Katz and Liebman to create a standardized treatment effect via the following steps: i) sign all outcomes in the domain so the hypothesized effects go in the same direction; ii) standardize each variable as a z-score by subtracting the control group mean and dividing by the control group standard deviation; and then iii) averaging these z-scores.
- 3) To control for multiple hypothesis testing with respect to the heterogeneity of treatment effects, we will follow the recommendations of Fink, McConnell and Vollmer (2010) and employ the Benjamini and Hochberg (1995) method to minimize the false non-discovery rate (FNR). We will also limit our examination of treatment effect heterogeneity to select outcomes identified in the hypotheses above.

5.6 Survey attrition

Depending on response rates and budget, the follow-up survey will potentially use more expensive methods to try and get a subsample of the individuals who can be obtained through the standard survey to respond. If this is done, all data will be probability-reweighted to reflect this.

Let A_i be an indicator of whether individual i attrits from the study by not responding to or being able to be contacted for a follow-up survey. We will first estimate whether attrition is related to treatment status by means of the following regression:

$$A_i = \beta_0 + \beta_1 T_i + X'_s \delta + \varepsilon_i$$

Where X_s are dummy variables for each randomization strata s . We will test $\beta_1 = 0$ to determine whether attrition from the survey is related to treatment status or not. Standard errors will be clustered at the market level when comparing individuals in treated vs non-treated markets, and at the individual level when comparing treated and control individuals within treated markets.

If treatment status is found not to significantly affect attrition at the 5 percent significance level, then all estimation will proceed without any adjustment for attrition. If attrition is found to be related to treatment status, we postulate that attrition will be higher for the control group. We will then employ two bounding approaches to test robustness to attrition:

- (i) Lee bounds: the group with lower attrition will have either the top or the bottom tail of responses trimming following the Lee method. For continuous outcomes robustness to assuming that the attrited observations were at the 95th, 90th, and 75th percentiles will be used for the lower bound, and 5th, 10th, 25th percentiles for the upper bound.
- (ii) Behaghel et al bounds: we will use the number of attempts it took to contact respondents to form bounds following the approach set out in their paper.

5.7 Missing data from item non-response

No imputation for missing data from item non-response at follow-up will be performed. Missing data on baseline variables will be dummied out of the ANCOVA specifications, as detailed above. We will check whether item non-response is correlated with treatment status following the same procedures as for survey attrition, and if it is, construct bounds for our treatment estimates that are robust to this.

5.8 Questions with Limited Variation

In order to limit noise caused by variables with minimal variation, questions for which 95 percent of observations have the same value within the relevant sample will be omitted from the construction of any aggregate index.

5.9 Inflation

Once multiple rounds of survey data are available, nominal values will be converted to real Kenyan Shillings using the monthly Kenya consumer price index from the National Statistics Bureau of Kenya.

6. Updated Power Calculations

Our original proposal estimated statistical power based on assumptions about the sample size, number of clusters, coefficient of variation, and intra-cluster correlation. Based on the baseline data and realized take-up rate for training, we have revisited these power calculations.

The proposal was based on 120 marketplaces, an assumed intra-cluster correlation of 0.05, a coefficient of variation of 0.8, a take-up rate of 75%, and an average sample size of 30 firms per market.

In practice we were reasonably close to these assumptions: we have 157 marketplaces, intra-cluster correlations ranging from 0.022 to 0.06 depending on the profits or sales measure we use, a coefficient of variation ranging from 0.77 to 1.33, a take-up rate of 77.7%, and an average sample size of 22 firms per market.

The table below then is based on the actual data. It calculates power for a simple comparison of means using only a single round of follow-up data. This is conservative, and ignores the gain in power from conditioning on baseline data (McKenzie, 2012). We calculate power for two scenarios: first, for comparing treated to control firms within the treated markets, assuming no spillovers. This results in a minimal detectable effect of a 12 to 20 percent increase in profits, which is below the base case of a 25% increase examined in our proposal, and much more powerful than existing studies (McKenzie and Woodruff, forthcoming).² The last column then is based on comparison of treated firms to control firms in control markets, accounting for the clustered randomization, and assuming 22 firms per market and 75 markets treated and 75 control as an approximation for the purpose of this calculation. Since the intra-cluster correlation

² E.g. to get the 12.1% increase for row 1, we use the Stata command `power twomeans 100 109.4, sd(77) n1(1173) n2(988)` to see this gives power 0.8, and then divide the effect of 9.4% by the take-up rate (77.7%) to get 12.1%.

is very low, this not result in much change in the minimal detectable effect from the no spillover case.

	Mean	Std. Dev.	C.V.	ICC	Minimal Detectable Effect	
					No Spillover	Spillovers
Census weekly profits	1117	856	0.77	0.06	12.1%	14.6%
Baseline monthly profits (truncated at 99th percentile)	4782	5275	1.10	0.031	17.2%	17.9%
Baseline weekly sales (truncated at 99th percentile)	6164	8177	1.33	0.022	20.7%	20.30%

Notes:

C.V. denotes coefficient of variation, ICC the intra-cluster correlation

Actual power will be greater once we allow for i) the use of randomization strata fixed effects (Bruhn and McKenzie, 2009); and ii) control for the lagged dependent variable (McKenzie, 2012). As a result, we are confident that the study has sufficient power as implemented to detect effects that are of economically meaningful size, and in particular, are of the size seen in other studies in the literature (McKenzie and Woodruff, forthcoming).

References:

Bauer, Susanne, Gerry Finnegan and Neliën Haspels (2004) *Gender and Entrepreneurship Together: GET Ahead for Women in Enterprise: Training Package and Resource Kit*, Bangkok, Berlin and Geneva, International Labour Office.

Behaghel, Luc, Bruno Crépon, Marc Gurgand, Thomas Le Barbanchon (2012) “Please Call Again: Correcting Non-Response Bias in Treatment Effect Models”, *IZA Discussion Paper no. 6751*.

Benjamini, Yoav and Yosef Hochberg (1995) —Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society Series B*, 57(1): 289-300.

Bruhn, Miriam and David McKenzie (2009) “In pursuit of balance: Randomization in practice in development field experiments”, *American Economic Journal: Applied Economics* 1(4): 200-32

De Mel, Suresh, David McKenzie and Christopher Woodruff (2014) Business Training and Female Enterprise Start-up, Growth, and Dynamics: Experimental evidence from Sri Lanka, *Journal of Development Economics*, 106: 199-210

Fink, Günther, Margaret McConnell and Sebastian Vollmer (2012) —Testing for Heterogeneous Treatment Effects in Experimental Data: False Discovery Risks and Correction Procedures, Mimeo. Harvard School of Public Health.

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

McKenzie, David and Christopher Woodruff (forthcoming) What are we learning from business training evaluations around the developing world? *World Bank Research Observer*.