Pre-Analysis Plan: Promoting entrepreneurship among migrants and host populations in Colombia[*]

Tommaso Crosta Dean Karlan Martin Valdivia

Abstract

The purpose of this study is to examine whether entrepreneurship training enhances the sustainability and growth of businesses run by migrants in Colombia. We will collaborate with iNNpulsa Colombia and USAID's Oportunidades sin Fronteras for the implementation of a randomized controlled trial to study the effects of business and soft-skills training on business outcomes as well as psychometric and household-level outcomes. By comparing traditional business training with a combined technical and soft skills approach, the study aims to determine the additional value of soft skills in fostering business success and improving the effectiveness of such programs. Additionally, through pre-treatment expectations we will explore how interventions can be better targeted to potential high-growth entrepreneurs. The findings will contribute to better targeting and cost-effectiveness of entrepreneurship support initiatives for migrants and other vulnerable populations.

January 2025

[*] Contact: Tommaso Crosta (tommaso.crosta@phd.unibocconi.it), Martin Valdivia (jvaldivi@grade.org.pe)

1 Introduction

Colombia is the primary destination for Venezuelan migrants, with nearly 2.9 million arriving as of December 2023 (Migración Colombia, 2023). Despite a regularization process by the Colombian government, Venezuelan migrants, particularly women, encounter significant discrimination and barriers to workforce entry (Amnesty International, 2020) and face costly challenges in validating educational and professional credentials (Guerrero Ble, 2023). While various programs aim to enhance migrant income generation, scalable and cost-effective solutions are crucial for addressing these issues effectively.

Many Venezuelan migrants turn to entrepreneurship due to labor market challenges and the flexibility it offers for child care (International Labour Organization [ILO], 2021). They are more inclined to start businesses than locals (Licheri et al., 2024) and foreign-owned firms in Colombia, often Venezuelan, are 10-20% more capitalized than local firms (Bahar et al., 2023). However, these firms have similar or slightly lower survival rates compared to local ones (Bahar et al., 2023; Licheri et al., 2024). As of 2020, 98.8% of Venezuelan entrepreneurs operated informally and faced issues such as lower education levels, fewer assets, and decreasing capital (Building Markets, 2023; Observatorio Proyecto Migración Venezuela, 2020).

In Colombia, various stakeholders, including government agencies, international organizations, and local NGOs, implement initiatives to integrate Venezuelan migrants into the entrepreneurial ecosystem. These programs often provide technical skills training, seed funding, and matchmaking events. However, implementers often face challenges including ineffective targeting methods, with some beneficiaries abandoning their ventures or continuing their migration, and a lack of robust, evidence-based soft-skills training. This research aims to address these issues by exploring cost-effective targeting strategies and evaluating the impact of incorporating soft-skills modules in these interventions.

Studies show that targeting policies for entrepreneurs effectively requires distinguishing between subsistence and transformational entrepreneurs, who respond differently to support (Schoar, 2010). Key factors influencing high-growth potential include prior ability (Fafchamps & Woodruff, 2017), community knowledge (Hussam et al., 2022), and financial support (McKenzie, 2017). This study aims to enhance the cost-effectiveness of business training programs by identifying factors that improve the success of Venezuelan entrepreneurs in Colombia. We will assess variables such as migration history, duration of displacement, dependents, prior business experience, and business composition to determine their impact on program outcomes.

Despite its benefits, evidence-based soft skills training is rarely included in interventions. Recent research indicates that soft skills are crucial for successful entrepreneurship, enhancing sales, profits, and overall business performance (Innovations for Poverty Action [IPA], 2023). IPA has identified such training as a Best Bet, a promising intervention likely to generate a high impact when scaled up. To address this gap, IPA Colombia, with support from the Conrad N. Hilton Foundation, has partnered with iNNpulsa and Corporación Mundial de la Mujer to pilot a comprehensive soft skills curriculum in 2024. This curriculum, based on a thorough evidence review and ongoing learning agenda, was first tested with 130 entrepreneurs and has been refined and made openly available by August 2024. The soft skills training curriculum is based on extensive evidence reviews and meta-analyses (McKenzie & Woodruff, 2017; McKenzie et al., 2023) and incorporates successful elements from programs such as Personal Initiative (Campos et al., 2017; Glaub et al., 2014; Ubfal et al., 2022), SEED (Chioda et al., 2021), SEE (Shankar et al., 2015), and StartUp! & ReachUp! (Alibhai et al., 2019). It emphasizes soft skills that enhance entrepreneurial performance, including increased sales, profits, and venture survival. Supported by the World Bank's Gender Innovation Lab and IPA's (Delavallade & Rouanet, 2020) agenda, the curriculum integrates best practices for teaching and program implementation, such as feedback mechanisms, collaborative environments, and trainer qualifications. This pilot represents the cutting edge of soft skills training in the field.

Our proposed study will be the first to rigorously evaluate the impact of a robust, evidence-based soft skills curriculum for migrant entrepreneurs in Colombia. It will assess whether the benefits of such training observed in other countries apply in Colombia, how different entrepreneur profiles (e.g., gender, migration status, sector) respond, and whether combining soft skills with technical training enhances business outcomes and social integration. The study will also explore how to adapt, scale, and replicate this training while ensuring quality.

2 Experimental design

2.1 Treatment Description

We consider two treatment arms and a control group. One treatment arm (T1) implies offering a state of the art business training program that can transmit key business practices associated with successful enterprises. The second treatment arm (T2) will combine the transmission of the same business practices as in T1 with the transmission of key soft skills through the open access soft skills curriculum developed by IPA and iNNpulsa. Both treatment arms will include a transfer of assets, tailored for the specific entrepreneur's needs. Comparing the treatment group T1 with the control group allows us to evaluate the impact of a traditional business training program and asset transfer over the success of the businesses run by the treated, plus some other key individual and household income indicators. Comparing T2 with the control group allows us to estimate the impact of the full treatment (hard + soft skills + asset transfer). The comparison of T1 and T2 allows us to establish the marginal contribution of the soft skills module.[1]

2.2 Random Assignment

Treatment assignment will be stratified and individually randomized, with an equal proportion of participants being assigned to each group (Control, T1, and T2). The total sample size will be around 1587 entrepreneurs, with 529 being assigned to each treatment group and 529 to the control group. Regarding the implementers, 13% (201) of the recruited sample corresponds to OSF, 36% (570) to iNNpulsa, and 51% (816) to Colombia Productiva.

2.3 Power Calculation

For the power calculations, in line with our expectations on sample size, we assume 1024 individuals treated across all implementers and locations, equally split into the two treatment arms, and 521 individuals in the control group. We assume a minimum detectable effect (MDE) of 0.25 for T1 and 0.3 for T2, a stratified randomization based on 7 variables and test the power from a standard regression of the outcome on the two treatment indicators and the strata variables. We average across 100 simulations and several seeds to obtain stable predictions. Notice that, since we plan to use a Bayesian model for more advanced analyses, namely heterogeneity, mediator effect and expectation analysis, we will be able to use prior distribution to regularize the estimation procedure, thus increasing the precision of estimates and, consequently, power (Iacovone et al., 2023).

Figure 1 shows the power for detecting T1 and T2 effects in relation with control group size. It shows that the group size we are working with allows us a statistical power higher than standard studies, but we consider such sample sizes will allow us to have enough power to develop the heterogeneity, mediator effect and expectation analysis proposed.

[1] We are coordinating with the implementers to organize the two modules (hard skills, soft skills) homogeneously throughout the training program, so that at any effective exposure, T2 beneficiaries will differentiate from T1 beneficiaries in that the former will include some training on soft skills.

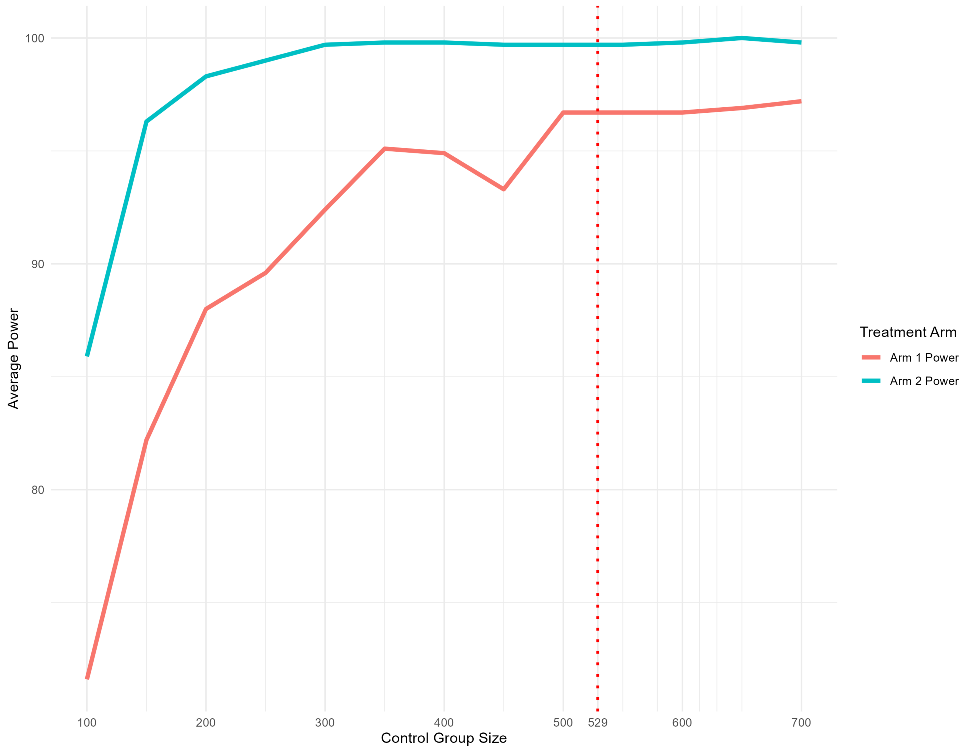

Figure 1: Average power curve for treatment arms

3 Empirical analysis

3.1 Average Treatment Effects estimation

Comparing the treatment group T1 with the control group allows us to evaluate the impact of a traditional business training program over the success of the businesses run by the treated, plus some other key individual and household income indicators. Comparing T2 with the control group allows us to estimate the impact of the full treatment (hard+soft skills). The comparison of T1 and T2 allows us to establish the marginal contribution of the soft skills module.

As perfect compliance cannot be guaranteed after treatment assignment, we will first estimate intent-to-treat (ITT) effects. To estimate the impact, we rely on the standard approach for stratified RCTs prescribed by Imbens and Rubin (2015), which does not rely on model-based assumptions of the data generating process, but only on design-based ones, namely SUTVA and random assignment by stratum. The drawback of this approach is that it does not use covariate adjustments to improve precision, especially past values of the outcome (McKenzie, 2012).

3.2 Bayesian analysis

In order to go beyond a simple analysis of ATEs, we will employ a Bayesian model[2], in line with recent advancements in the literature (Imbens and Rubin, 2015, Iacovone et al., 2023) to model potential outcomes, in order to:

1. introduce covariate adjustments into the analysis and improve precision of the effects through the use of priors elicited by aggregating respondents' beliefs
2. estimate heterogeneous treatment effects on a pre-defined set of respondents' characteristics
3. model the potential outcomes of respondents as in Imbens and Rubin (2015), in order to predict Conditional Individual Treatment Effects (CITEs) and Quantile Treatment Effects (QTEs)
4. compare predicted CITEs, with respondents' own beliefs about their own future outcomes
5. use a hierarchical model to investigate the mediating role of business practices increasing the revenues, profits and other business outcomes

3.2.1 Bayesian model specifications

3.2.1.1 Continuous Outcomes

3.2.1.1.1 Symmetric Variables

For continuous variables with approximately **symmetric** distribution (and no discrete mass around zero) we can use a standard hierarchical Normal model, i.e.:

[2] in line with Gelman (2006), although not mentioned, we will use half-Cauchy priors for scale parameters in all models.

$$Y|(\alpha_s)_{s=1}^S, \theta, \beta, \Sigma_y, X \sim \mathcal{MN}\left(\sum_{s=1}^S \alpha_s D_s + \theta T + \boldsymbol{\beta}(X - \bar{X}), \Sigma_y\right)$$

$$\theta \sim \hat{f}(m(1) - m(0))$$

$$\alpha_s \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha \sim \hat{g}(m(0))$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(\mathbf{0}, \Sigma_\beta)$$

Where $\hat{f}$ and $\hat{g}$ are elicited distributions and $s \in \{1, ..., S\}$ are the strata and D_s is the stratum dummy and $m_i(k) = \mathbb{E}(Y_i|(\alpha_s)_{s=1}^S, \theta, \beta, \sigma_y, X_i, T_i = k)$

Eliciting strategy:

(i) elicit $(m_i(0), m_i(1))_{i=1}^N$ in the sample

(ii) compute the sample distribution of $\hat{\theta} = m(1) - m(0)$ and $\hat{\alpha} = m(0)$

(iii) fit prior $\hat{f}(m(1) - m(0))$ and $\hat{g}(m(0))$

3.2.1.1.2 Skewed Variables

For example, sales or costs.

$$Y_i|\alpha_{s(i)}, \theta, \beta, \sigma_y, X_i \overset{\text{ind}}{\sim} log\mathcal{N}\left(\sum_{s=1}^{S} \alpha_s D_{s(i)} + \theta T_i + (X_i - \bar{X})'\boldsymbol{\beta}, \sigma_y\right)$$

$$\theta \sim \hat{f}(log(m(1)) - log(m(0)))$$

$$\alpha_s \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha \sim \mathcal{N}(\overline{log(y(0))}, \sigma_0)$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(\mathbf{0}, \Sigma_\beta)$$

Where $m(k) = \mathbb{E}(Y_i|(\alpha_s)_{s=1}^S, \theta, \beta, \sigma_y, X_i, T_i = k)$. Crucially, with a log-normal distribution, we have:

$$\mathbb{E}(Y_i|(\alpha_s)_{s=1}^S, \theta, \beta, \sigma_y, X_i) = exp\left\{\sum_{s=1}^{S} \alpha_s D_{s(i)} + \theta T_i + (X_i - \bar{X})'\boldsymbol{\beta} + \frac{\sigma_y}{2}\right\}$$

Eliciting strategy:

(i) elicit $(m_i(0), m_i(1))_{i=1}^N$ in the sample

(ii) compute $(log(m_i(0)), log(m_i(1)))_{i=1}^N$

(iii) compute the sample distribution of $\hat{\theta} = log(m(1)) - log(m(0))$

(iv) fit prior $\hat{f}(log(m(1)) - log(m(0)))$

Notice that, given the distributional assumption we cannot identify a prior for α using elicitation data.

3.2.1.1.3 Zero-inflated Skewed Variables

Variables such as sales, might present a non negligible probability mass around zero, in this case the previous model is not feasible and needs to be "inflated" by a mass of zeros. Such model is called zero-inflated Hurdle model.

$$Y_i|p_i,\mu,\sigma_y = \begin{cases} 0 & \overset{\text{ind}}{\sim} p_i, \qquad \text{if } Y_i = 0 \\ log\mathcal{N}(\mu,\sigma_y) & \overset{\text{ind}}{\sim} 1-p_i, \quad \text{if } Y_i > 0 \end{cases}$$

$$\mu_i = \sum_{s=1}^{S} \alpha_{s,I} D_{s(i)} + \theta_I T_i + (X_i - \bar{X})'\boldsymbol{\beta_I} \qquad \text{[Intensive Margin]}$$

$$p_i \overset{\text{ind}}{\sim} \mathcal{B}ernoulli(\pi_i)$$

$$\pi_i = \sum_{s=1}^{S} \alpha_{s,E} D_{s(i)} + \theta_E T_i + (X_i - \bar{X})'\boldsymbol{\beta_E} \qquad \text{[Extensive Margin]}$$

$$\theta_I \sim \hat{f}_I(log(m(1)) - log(m(0)))$$

$$\theta_E \sim \hat{f}_E\left(log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$$

$$\alpha_{s,I} \sim \mathcal{N}(\alpha,\sigma_\alpha)$$

$$\alpha_I \sim \mathcal{N}(\overline{log(y(0))},\sigma_0)$$

$$\alpha_{s,E} \sim \mathcal{N}(\alpha,\sigma_\alpha)$$

$$\alpha_E \sim \hat{g}\left(log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(\mathbf{0},\Sigma_\beta)$$

Elicitation strategy:

(A) elicit $\hat{f}_E$

- (i) elicit $(\pi_i(0),\pi_i(1))_{i=1}^{N}$ in the sample
- (ii) compute $\left(log\left(\frac{\pi_i(0)}{1-\pi_i(0)}\right), log\left(\frac{\pi_i(1)}{1-\pi_i(1)}\right)\right)_{i=1}^{N}$

(iii) compute the sample distribution of $\hat{\theta} = log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)$ and $\hat{\alpha} = log\left(\frac{\pi(0)}{1-\pi(0)}\right)$

(iv) fit priors $\hat{f}_E\left(log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$ and $\hat{g}\left(log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$

(B) elicit $\hat{f}_I$

(i) elicit $(m_i(0), m_i(1))_{i=1}^N$ in the sample

(ii) compute $(log(m_i(0)), log(m_i(1)))_{i=1}^N$

(iii) compute the sample distribution of $\hat{\theta}_I = log(m(1)) - log(m(0))$

(iv) fit prior $\hat{f}_I(log(m(1)) - log(m(0)))$

3.2.1.1.4 Skewed Real Variables

For example, sales or costs.

$$Y_i|\alpha_{s(i)}, \theta, \beta, \sigma_y, X_i \overset{\text{ind}}{\sim} \mathcal{G}umbell\left(\sum_{s=1}^{S} \alpha_s D_{s(i)} + \theta T_i + (X_i - \bar{X})'\boldsymbol{\beta}, \sigma_y\right)$$

$$\theta \sim \hat{f}(m(1) - m(0))$$

$$\alpha_s \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha \sim \mathcal{N}(\overline{log(y(0))}, \sigma_0)$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(\mathbf{0}, \Sigma_\beta)$$

Where $m(k) = \mathbb{E}(Y_i|(\alpha_s)_{s=1}^S, \theta, \beta, \sigma_y, X_i, T_i = k)$. Crucially, with a log-normal distribution, we have:

$$\mathbb{E}(Y_i|(\alpha_s)_{s=1}^S, \theta, \beta, \sigma_y, X_i) = \sum_{s=1}^{S} \alpha_s D_{s(i)} + \theta T_i + (X_i - \bar{X})'\boldsymbol{\beta} + \sigma_y \gamma^{3}$$

Eliciting strategy:

(i) elicit $(m_i(0), m_i(1))_{i=1}^N$ in the sample

(ii) compute the sample distribution of $\hat{\theta} = m(1) - m(0)$

(iii) fit prior $\hat{f}(m(1) - m(0))$

Notice that, given the distributional assumption we cannot identify a prior for α using elicitation data.

3.2.1.1.5 Zero-inflated Skewed Variables

Variables such as sales, might present a non negligible probability mass around zero, in this case the previous model is not feasible and needs to be "inflated" by a mass of zeros. Such model is called zero-inflated Hurdle model.

[3] γ is the Euler-Mascheroni constant

$$Y_i|p_i,\mu,\sigma_y = \begin{cases} 0 & \overset{\text{ind}}{\sim} p_i, \quad \text{if } Y_i = 0 \\ \mathcal{G}umbell(\mu,\sigma_y) & \overset{\text{ind}}{\sim} 1-p_i, \quad \text{if } Y_i > 0 \end{cases}$$

$$\mu_i = \sum_{s=1}^{S} \alpha_{s,I} D_{s(i)} + \theta_I T_i + (X_i - \bar{X})'\boldsymbol{\beta_I} \qquad \text{[Intensive Margin]}$$

$$p_i \overset{\text{ind}}{\sim} \mathcal{B}ernoulli(\pi_i)$$

$$\pi_i = \sum_{s=1}^{S} \alpha_{s,E} D_{s(i)} + \theta_E T_i + (X_i - \bar{X})'\boldsymbol{\beta_E} \qquad \text{[Extensive Margin]}$$

$$\theta_I \sim \hat{f}_I(m(1) - m(0))$$

$$\theta_E \sim \hat{f}_E\left(log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$$

$$\alpha_{s,I} \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha_I \sim \mathcal{N}(\overline{log(y(0))}, \sigma_0)$$

$$\alpha_{s,E} \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha_E \sim \hat{g}\left(log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(\mathbf{0}, \Sigma_\beta)$$

Elicitation strategy:

(A) elicit $\hat{f}_E$

(i) elicit $(m_i(0), m_i(1))_{i=1}^{N}$ in the sample

(ii) compute the sample distribution of $\hat{\theta} = m(1) - m(0)$

(iii) fit prior $\hat{f}(m(1) - m(0))$

(B) elicit $\hat{f}_I$

(i) elicit $(m_i(0), m_i(1))_{i=1}^N$ in the sample

(ii) compute $(log(m_i(0)), log(m_i(1)))_{i=1}^N$

(iii) compute the sample distribution of $\hat{\theta}_I = log(m(1)) - log(m(0))$

(iv) fit prior $\hat{f}_I(log(m(1)) - log(m(0)))$

3.2.1.2 Count Variables

3.2.1.2.1 Standard Poisson Model

$$Y_i|\mu_i \overset{\text{ind}}{\sim} \mathcal{P}oisson(\mu_i)$$

$$log(\mu_i) = \sum_{s=1}^{S} \alpha_s D_{i,s} + \theta T_i + (X_i - \bar{X})'\boldsymbol{\beta}$$

$$\alpha_s \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha \sim \hat{g}\left(log(\mu(0))\right)$$

$$\theta \sim \hat{f}\left(log(\mu(1)) - log(\mu(0))\right)$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(0, \Sigma_\beta)$$

Where $\mu_i(k) = log(\mu_i)\Big|_{T_i=k}, \quad k = \{0, 1\}$.

Keeping in mind that $\mathbb{E}(Y_i|\mu_i) = \mu_i$, i.e. μ_i represents the average duration, the eliciting strategy is:

(i) elicit $(\mu_i(0), \mu_i(1))_{i=1}^N$ in the sample

(ii) compute $(log(\mu_i(0)), log(\mu_i(1)))_{i=1}^N$

(iii) compute the sample distribution of $\hat{\theta} = log(\mu(1)) - log(\mu(0))$ and $\hat{\alpha} = log(\mu(0))$

(iv) fit priors $\hat{f}(log(\mu(1)) - log(\mu(0)))$ and $\hat{g}(log(\mu(0)))$

3.2.1.2.2 Zero-inflated Poisson model

$$Y_i|p_i,\mu,\sigma_y = \begin{cases} 0 & \overset{\text{ind}}{\sim} \pi_i, \quad \text{if } Y_i = 0 \\ \mathcal{P}oisson(\mu_i) & \overset{\text{ind}}{\sim} 1-\pi_i, \quad \text{if } Y_i > 0 \end{cases}$$

$$log(\mu_i) = \sum_{s=1}^{S} \alpha_{s,I} D_{s(i)} + \theta_I T_i + (X_i - \bar{X})'\boldsymbol{\beta_I} \qquad \text{[Intensive Margin]}$$

$$p_i \overset{\text{ind}}{\sim} \mathcal{B}ernoulli(\pi_i)$$

$$\pi_i = \sum_{s=1}^{S} \alpha_{s,E} D_{s(i)} + \theta_E T_i + (X_i - \bar{X})'\boldsymbol{\beta_E} \qquad \text{[Extensive Margin]}$$

$$\theta_I \sim \hat{f}_I \left(log(\mu(1)) - log(\mu(0))\right)$$

$$\theta_E \sim \hat{f}_E \left(log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$$

$$\alpha_{s,I} \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha \sim \hat{g}_I \left(log(\mu(0))\right)$$

$$\alpha_{s,E} \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha_E \sim \hat{g}_E \left(log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(0, \Sigma_\beta)$$

Elicitation strategy:

(A) elicit $\hat{f}_E$

(i) elicit $(\pi_i(0), \pi_i(1)\mu(0)_{i,E}, \mu(1)_{i,E})_{i=1}^N$ in the sample

(ii) compute $\left(log\left(\frac{\pi_i(0)}{1-\pi_i(0)}\right), log\left(\frac{\pi_i(1)}{1-\pi_i(1)}\right)\right)_{i=1}^N$

    (iii) compute the sample distribution of $\hat{\theta} = log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)$ and $\hat{\alpha} = log\left(\frac{\pi_(0)}{1-\pi(0)}\right)$

    (iv) fit priors $\hat{f}_E\left(log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$ and $\hat{g}_I\left(log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$

(B) elicit $\hat{f}_I$

    (i) elicit $(\mu_i(0), \mu_i(1))_{i=1}^N$ in the sample

    (ii) compute $(log(\mu_i(0)), log(\mu_i(1)))_{i=1}^N$

    (iii) compute the sample distribution of $\hat{\theta} = log(\mu(1)) - log(\mu(0))$ and $\hat{\alpha} = log(\mu(0))$

    (iv) fit priors $\hat{f}(log(\mu(1)) - log(\mu(0)))$ and $\hat{g}_E(log(\mu(0)))$

3.2.1.3 Dichotomous Variables

$$Y_i|\pi \overset{\text{ind}}{\sim} Bernoulli(\pi_i)$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{s=1}^{S} \alpha_s D_{i,s} + \theta T_i + (X_i - \bar{X})'\boldsymbol{\beta}$$

$$\alpha_s \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha \sim \hat{g}\left(\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$$

$$\theta \sim \hat{f}\left(\left(log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)\right)$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(0, \Sigma_\beta)$$

Keeping in mind that $\mathbb{E}(Y_i|\pi_i) = \pi_i$, i.e. π_i represents the expected probability of $Y_i = 1$. Let $\pi_i(k) = \pi_i\Big|_{T_i=k}, \quad k = \{0, 1\}$. The eliciting strategy is:

(i) elicit $(\pi_i(0), \pi_i(1))_{i=1}^N$ in the sample

(ii) compute $\left(log\left(\frac{\pi_i(0)}{1-\pi_i(0)}\right), log\left(\frac{\pi_i(1)}{1-\pi_i(1)}\right)\right)_{i=1}^N$

(iii) compute the sample distribution of $\hat{\theta} = log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)$ and $\hat{\alpha} = log\left(\frac{\pi(0)}{1-\pi(0)}\right)$

(iv) fit priors $\hat{f}\left(log\left(\frac{\pi(1)}{1-\pi(1)}\right) - log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$ and $\hat{g}\left(log\left(\frac{\pi(0)}{1-\pi(0)}\right)\right)$

3.2.2 Mediator analysis

Considering the simplest normal model, given an outcome Y, a matrix of mediator outcomes M and control variables X, then the model specification would be:

$$Y|(\alpha_{s,y})_{s=1}^{S}, \theta, \beta, \Sigma_y, X, M \sim \mathcal{MN}\left(\sum_{s=1}^{S} \alpha_{s,y} D_s + \theta_D T + \boldsymbol{\beta}_y (X_y - \bar{X}_y) + \gamma M, \Sigma_y\right)$$

$$M|(\alpha_{s,m})_{s=1}^{S}, \theta, \beta, \Sigma_m, X, M \sim \mathcal{MN}\left(\sum_{s=1}^{S} \alpha_{s,m} D_s + \theta_M T + \boldsymbol{\beta}_m (X_m - \bar{X}_m), \Sigma_m\right)$$

$$\theta_D \sim \hat{f}(m(1) - m(0))$$

$$\gamma \sim \mathcal{N}(0, 25)$$

$$\theta_M \sim \mathcal{N}(0, 25)$$

$$\alpha_s \sim \mathcal{N}(\alpha, \sigma_\alpha)$$

$$\alpha \sim \hat{g}(m(0))$$

$$\boldsymbol{\beta} \sim \mathcal{MN}(\mathbf{0}, \Sigma_\beta)$$

Where θ_M is the ATEs on M, $\theta_I = \gamma * \theta_M$ is the indirect ATEs on Y, mediated by M and θ_D is the direct ATE. Then, the overall ATE of T on Y is $\theta_T = \theta_D + \theta_I$.

3.2.3 Heterogeneity Analysis

We plan to study heterogeneity, using the same Bayesian models outlined above, to simulate the missing potential outcome to estimate, for a given group $\mathcal{G}$ defined in $supp(X)$:

$$\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}(Y(1)-Y(0))$$

Where, for $s \in \{0,1\}$

$$Y_n(s) = \begin{cases} Y_{n,obs} \quad if \;\; Y_{n,obs} = Y_n(s) \\ Y_{n,sampled} \sim f(Y_n(s)|X) \end{cases}$$

We plan to model a correlation between $(Y(1), Y(0))$, though this will not be straightforward when

The variables we plan to analyze are: gender, migration status, dimension of the business (by sales, number of employees), age of business, educational level, household characteristics (e.g. household head, family composition), internally-displaced population, host community entrepreneurs, amount of time residing in Colombia, individual psychometric variables such as perceived stress, personal initiative, grit and locus of control.

3.2.4 Prior estimation

We developed a novel survey instrument to elicit participants' prior probabilistic beliefs about their conditional potential outcomes $Y(1), Y(0)|X$. We will follow the methodology of Gosling (2018), recently applied in Iacovone et al. (2023), in order to estimate prior distribution from the elicited probabilistic expectations about several primary outcomes such as business sales, business survival and number of employees.

3.2.5 Instrumental Variable Estimation

In order to go beyond ITT analysis, we use the approach of Imbens and Rubin (1997), Imbens and Rubin (2015) and Lee et al. (2022). We assume two latent compliance groups:

- **Compliers (C)**: Take the treatment when assigned ($Z = 1$), do not take it when assigned to control ($Z = 0$).
- **Never-Takers (N)**: Never take the treatment ($W = 0$), regardless of assignment.

Let π_C be the probability of being a **complier**:

$$\pi_C = P(\text{Complier}), \quad 1 - \pi_C = P(\text{Never-Taker}).$$

The received treatment W depends on the assigned treatment Z:

$$W_n = \begin{cases} 1, & \text{if } Z_n = 1 \text{ and individual is a Complier,} \\ 0, & \text{otherwise.} \end{cases}$$

The observed outcome Y_n follows a normal distribution:

$$\begin{aligned} Y_n | (Z_n = 1, W_n = 1, C) &\sim \mathcal{N}(\mu_{C1}, \sigma_Y^2), \quad \text{(Complier under Treatment)} \\ Y_n | (Z_n = 0, W_n = 0, C) &\sim \mathcal{N}(\mu_{C0}, \sigma_Y^2), \quad \text{(Complier under Control)} \\ Y_n | (Z_n = 1, W_n = 0, N) &\sim \mathcal{N}(\mu_N, \sigma_Y^2), \quad \text{(Never-Taker under any } Z) \end{aligned}$$

For each individual n, the likelihood depends on the observed assignment Z_n and treatment received W_n:

- **If assigned to treatment** $(Z_n = 1)$:

$$\begin{aligned} P(Y_n | Z_n = 1, W_n = 1) &= \pi_C \cdot \mathcal{N}(Y_n | \mu_{C1}, \sigma_Y^2), \\ P(Y_n | Z_n = 1, W_n = 0) &= (1 - \pi_C) \cdot \mathcal{N}(Y_n | \mu_N, \sigma_Y^2). \end{aligned}$$

- **If assigned to control** $(Z_n = 0, W_n = 0)$:
  The compliance status is unknown, so we use a mixture model:

$$P(Y_n | Z_n = 0, W_n = 0) = \pi_C \cdot \mathcal{N}(Y_n | \mu_{C0}, \sigma_Y^2) + (1 - \pi_C) \cdot \mathcal{N}(Y_n | \mu_N, \sigma_Y^2).$$

The total log-likelihood is:

$$\log P(Y|Z,W) = \sum_{n=1}^{N} \begin{cases} \log \pi_C + \log \mathcal{N}(Y_n|\mu_{C1}, \sigma_Y^2), & \text{if } Z_n = 1, W_n = 1, \\ \log(1-\pi_C) + \log \mathcal{N}(Y_n|\mu_N, \sigma_Y^2), & \text{if } Z_n = 1, W_n = 0, \\ \log\left[\pi_C \cdot \mathcal{N}(Y_n|\mu_{C0}, \sigma_Y^2) + (1-\pi_C) \cdot \mathcal{N}(Y_n|\mu_N, \sigma_Y^2)\right], & \text{if } Z_n = 0, W_n = 0. \end{cases}$$

We specify the following priors:

$$\begin{aligned} \pi_C &\sim \text{Uniform}(0,1), \\ \mu_{C1}, \mu_{C0}, \mu_N &\sim \mathcal{N}(X\beta, \sigma_\mu)^{4}, \\ \beta &\sim \mathcal{N}(0,10), \\ \sigma_Y, \sigma_\mu &\sim \text{Half-Cauchy}(0,5). \end{aligned}$$

The **Complier Average Causal Effect (CACE)** is the treatment effect for compliers:

$$\text{CACE} = \mu_{C1} - \mu_{C0}.$$

We will also be able to perform sensitivity checks regarding potential violations of the exclusion restriction, i.e. the potential outcomes for never-takers, who would not receive the treatments even if assigned to them, are unaffected by the assignment variable Z.

3.3 Text analysis of in-kind transfer items

Both treatment arms will entail a final in-kind transfers of assets to the participants. Such transfer will be of similar value across participants, but crucially will differ in its content, since it will be tailored to the needs of the entrepreneuer and planned between them and an implementer consultant. We plan to standardize the procurement forms so that we can use them as input for text analysis to identify what items correlated with higher treatment effects.

[4] Where X is a matrix of exogenous covariates.

3.4 Network Analysis

Although we were not able to collect direct network data to investigate potential informational spillover among respondents, we will be able to proxy social networks using information on whether participants attended the business classes in the the same group and also geographical proximity of participants' business or household location.

3.5 Cost Effectiveness Analysis

We plan to gather precise cost data from the implementing partners in order to standardize our treatment effects by program cost. Moreover, in line with Iacovone et al. (2023), and using the estimated Bayesian model from the afore mentioned analyses, we plan to estimate the probability of attaining the minimum viable threshold for scaling up the policy under different targeting scenarios.

4 Outcomes of the Analysis

4.1 Impact on respondents' business outcomes

1. Sales
2. Profits
3. Business survival
4. Capital investment
   (a) Purchases of business assets
   (b) Stock of business assets (K)
5. Number of workers (L)
6. Total Factor Productivity (TFP)
   - As in Bruhn et al. (2018), we obtain TFP by first estimating the following Cobb-Douglas production function:[5]

$$ln(Sales_{ipt}) = \lambda + \alpha_l ln(L_{ipt}) + \alpha_k ln(K_{ipt}) + \varepsilon_{ipt} \tag{1}$$

[5] We plan to explore alternative methods for estimating the production function such as the one used in Atkin et al. (2017), which uses past labor as an instrument for current labor, but at the cost of losing one period for the estimation. We are also planning to use a Bayesian model to be able to be more flexible in the estimation procedure within our mediating analysis

- And we define total factor productivity as:

$$TFP_{ipt} = ln(Sales_{ipt}) - \hat{\alpha}_l ln(L_{ipt}) - \hat{\alpha}_k ln(K_{ipt}) \tag{2}$$

4.2 Mechanisms for impact on businesses

1. Business practices index[6]
   - We will build the business practices index as an aggregate z-score of the 3 sub-indices. Each sub-index will in itself be an aggregate z-score of standardized dummies representing each item in the category.

   (a) Marketing practices

   (b) Costing and record-keeping practices

   (c) Financial planning practices
2. Business loans

   (a) Business has an active loan

   (b) Business has an active loan from a formal institution

   (c) Total value of active business loans
3. Business has a commercial partnership

   (a) Vertical (distribution, inputs, other)

   (b) Horizontal (joint sales/production, asset sharing, other)
4. Formalization

   (a) Tax registration

   (b) Chamber of commerce registration

   (c) Has juridical personhood
5. Firm changes main products
6. Hours dedicated to business

[6] Our business practices index is based on the practices highlighted by McKenzie and Woodruff (2017) as important for business growth.

4.3 Impact on respondents' soft skills (Follow-up soft-skills to be determined)

For each soft skill, we will form an aggregate z-score from standardized dummies indicating that the respondent agrees or strongly agrees (disgree or strongly disagree in the case of reverse-scored items) with each statement.

1. Personal Initiative
2. Grit-S
3. Negotiation
   (a) With clients
   (b) With providers
   (c) With workers

4.4 Impact on respondents' labour supply

1. Respondent has a job
2. Respondent has dependent work
3. Income from all occupations
4. Weekly hours worked

4.5 Impact on respondents' mental health

1. Perceived Stress Scale

4.6 Impact on the respondents' households

1. Household income
   - Household total income is calculated as the sum of:[7]
     + Total HH business profits
     + Total HH waged income
     + Total HH income from other sources

[7] Respondents are given the chance correct our initial estimate if they believe it necessary

2. Household labor supply
    - Total household labor supply is the sum of:
        + Hours worked in primary occupation by main income earners
        + Hours worked in HH businesses by main income earners
        + Hours worked in secondary occupation by respondent
3. Household assets
    (a) Durable goods index (Aggregate z-score? PCA? Another method?)
    (b) Total household savings
4. Household Debt
    (a) Household has an active loan
    (b) Household has an active loan from a formal institution
    (c) Total value of active household loans

References

Alibhai, S., Buehren, N., Frese, M., Goldstein, M., Papineni, S., & Wolf, K. (2019). *Full Esteem Ahead? Mindset-Oriented Business Training in Ethiopia*, World Bank. https://doi.org/10.1596/1813-9450-8892

Amnesty International. (2020). *Unprotected: Gender-Based Violence Against Venezuelan Refugee Women in Colombia and Peru*. https://www.amnesty.org/en/wp-content/uploads/2022/07/AMR0156752022ENGLISH.pdf

Atkin, D., Khandelwal, A. K., & Osman, A. (2017). Exporting and firm performance: Evidence from a randomized experiment. *The quarterly journal of economics*, *132*(2), 551–615.

Bahar, D., Cowgill, B., & Guzman, J. (2023). Refugee entrepreneurship: The case of venezuelans in colombia. *AEA Papers and Proceedings*, *113*, 352–356.

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

Building Markets. (2023). *Empowering inclusivity: A roadmap for unlocking the economic potential of MSMEs in Colombia*. https://static1.squarespace.com/

static/66708a0c236025663ebcd96f/t/66aa3ecf0ac22e33aaa2f192/1722433238929/Empowering_Inclusivity_2023.pdf

Campos, F., Frese, M., Goldstein, M., Iacovone, L., Johnson, H. C., McKenzie, D., & Mensmann, M. (2017). Teaching personal initiative beats traditional training in boosting small business in west africa. *Science*, *357*(6357), 1287–1290.

Chioda, L., Contreras-Loya, D., Gertler, P., & Carney, D. (2021). *Making entrepreneurs: Returns to training youth in hard versus soft business skills* (tech. rep.). National Bureau of Economic Research.

Delavallade, C. A., & Rouanet, L. M. (2020). *Unpacking socio-emotional skills for women's economic empowerment* (No. 154431). World Bank Group. Retrieved January 2, 2025, from https://documentos.bancomundial.org/es/publication/documents-reports/documentdetail/333911605591192864/Unpacking-Socio-Emotional-Skills-for-Women-s-Economic-Empowerment

Fafchamps, M., & Woodruff, C. (2017). Identifying gazelles: Expert panels vs. surveys as a means to identify firms with rapid growth potential. *The World Bank Economic Review*, *31*(3), 670–686.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Glaub, M. E., Frese, M., Fischer, S., & Hoppe, M. (2014). Increasing personal initiative in small business managers or owners leads to entrepreneurial success: A theory-based controlled randomized field intervention for evidence-basedmanagement. *Academy of Management Learning & Education*, *13*(3), 354–379.

Gosling, J. P. (2018). SHELF: The Sheffield Elicitation Framework. In L. C. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation* (pp. 61–93, Vol. 261). Springer, Cham. https://doi.org/10.1007/978-3-319-65052-4_4

Guerrero Ble, M. (2023). *A Forgotten Response and An Uncertain Future: Venezuelans' Economic Inclusion in Colombia*. Refugees International. https://d3jwam0i5codb7.cloudfront.net/wp-content/uploads/2023/11/Colombia-Report-November-2023.pdf

Hussam, R., Rigol, N., & Roth, B. N. (2022). Targeting high ability entrepreneurs using community information: Mechanism design in the field. *American Economic Review*, *112*(3), 861–98.

Iacovone, L., McKenzie, D., & Meager, R. (2023, January). *Bayesian impact evaluation with informative priors: An application to a colombian management and export improvement program* (Policy Research Working Paper No. 10274). World Bank Development Research Group. https://documents.worldbank.org/en/

publication / documents - reports / documentdetail / 099550001202312667 / pdf / IDU0e1840a110175404c830b06e04d8d87f5517b.pdf

Imbens, G. W., & Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, *25*(1), 305–327.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Innovations for Poverty Action. (2023). *Best bets: Emerging oportunities for impact at scale*. https://poverty-action.org/sites/default/files/2023-11/IPA-Best-Bets-Report-2023.pdf

International Labour Organization. (2021). *Migration from venezuela: Opportunities for latin america and the caribbean*. https://www.ilo.org/sites/default/files/wcmsp5/groups/public/%40americas/%40ro-lima/documents/publication/wcms_775183.pdf

Lee, J., Feller, A., & Rabe-Hesketh, S. (2022). Instrumental variables analysis of randomized experiments with one-sided noncompliance [Accessed: 2025-01-17]. https://mc-stan.org/learn-stan/case-studies/cace_one-sided.html

Licheri, D., Arenas-Ortiz, C., Henao Aristizábal, P., Hernández León, A., Rojas Calle, J. D., & Silupú Peñaranda, R. (2024). *Estudio de impacto fiscal de la migración venezolana en Colombia: Realidad vs. potencial*. https://respuestavenezolanos.iom.int/sites/g/files/tmzbdl526/files/documents/2024-04/informe.pdf

McKenzie, D. (2012). Beyond baseline and follow-up: The case for more t in experiments. *Journal of Development Economics*, *99*(2), 210–221.

McKenzie, D. (2017). Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition. *American Economic Review*, *107*(8), 2278–2307.

McKenzie, D., & Woodruff, C. (2017). Business practices in small firms in developing countries. *Management Science*, *63*(9), 2967–2981.

McKenzie, D., Woodruff, C., Bjorvatn, K., Bruhn, M., Cai, J., Gonzalez-Uribe, J., Quinn, S., Sonobe, T., & Valdivia, M. (2023). Training entrepreneurs. *VoxDevLit*, *1*(3).

Migración Colombia. (2023). *Radiografía de migrantes venezolanas(os) en Colombia: Corte 31 de diciembre de 2023*. https://www.migracioncolombia.gov.co/sites/unidad-administrativa-especial-migracion-colombia/content/files/001127/56343_informe-distribucion-migrantes-venezolanos-diciembre-2023-ejecutivo.pdf

Observatorio Proyecto Migración Venezuela. (2020). Emprendimiento de los migrantes venezolanos en colombia. *Revista Semana*.

Schoar, A. (2010). The divide between subsistence and transformational entrepreneurship. *Innovation policy and the economy*, *10*(1), 57–81.

Shankar, A. V., Onyura, M., & Alderman, J. (2015). Agency-based empowerment training enhances sales capacity of female energy entrepreneurs in kenya. *Journal of health communication*, *20*(sup1), 67–75.

Ubfal, D., Arraiz, I., Beuermann, D. W., Frese, M., Maffioli, A., & Verch, D. (2022). The impact of soft-skills training for entrepreneurs in jamaica. *World Development*, *152*, 105787.