THEORY-BASED REASONING AND PERFORMANCE[1]

ARNALDO CAMUFFO

ALFONSO GAMBARDELLA

DIEGO JANNACE

January 2026

Abstract

While theory-based decision-making is widely discussed in economic literature, its practical impact on decision performance remains largely untested. We develop a dynamic model of decision-making under uncertainty, showing that causal theories raise economic performance. We provide the first large-scale empirical evidence of this phenomenon by conducting eight field randomized controlled trials with 1,556 early-stage startups across eight countries, comparing theory-and-evidence-based training to evidence-only and control conditions. We estimate that theory-trained entrepreneurs achieve significantly higher growth in cumulative sales and employee productivity relative to both counterfactual groups. Additional analyses through specific measures of theory quality and experimentation reinforce that the former drives these performance gains, whereas experimentation alone offers no significant short-term benefits. Our findings establish theories as a critical determinant of economic performance and advocate embedding causal theorizing alongside experimentation in decision-making.

Preliminary draft, please do not quote or circulate without authors' permission.

[1]The authors gratefully acknowledge financial support from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme through Grant No.101021061.

INTRODUCTION

The importance of theory-based reasoning is well-established in economics and decision science (Gilboa and Samuelson, 2012; Karni, 2022), but less so in the practice of many professions. Theory-based reasoning is especially relevant in highly uncertain environments and for novel problems, where economic actors lack concrete experience (Gilboa, Minardi, and Samuelson, 2020). On many occasions, decisions are reactive, based on trial and error, and decision-makers often fail to distinguish between correlation and causation, or to think counterfactually. This generates costly misinterpretations of data and decision errors, leading to losses in decision efficiency and performance. Conversely, theory-based reasoning involves building a structured causal model (i.e., the theory) that maps antecedents and relevant contextual factors into a subjective distribution of possible outcomes.

Causal theories can benefit economic actors in three ways. First, they can define which variables causally affect their outcomes of interest, enabling them to focus on the activation of these causal antecedents to generate the desired outcome. Second, they can focus the acquisition of information on whether it is plausible to generate these causal antecedents, or on whether they causally affect the outcome of interest, reducing wasted effort on uninformative or irrelevant signals. Third, solving decision problems using forward causal theories implies that decision-makers – instead of merely asking "why" ex-post – focus on asking "what will happen if we do this?", emphasizing the *estimation of effects of causes* over *causes of effects* (Gelman and Imbens, 2013).

Despite the importance of theory-based reasoning, there is limited evidence on how it impacts real-world decisions and performance. We focus on the context of early-stage entrepreneurs who have to decide how to launch or scale their startups. This is an ideal testbed for several reasons. First, it is an uncertain context, as they have to assess the viability of new product-market fits. Second, founders are well-identified individuals, so that we can focus on who makes the decision. Third, it is not difficult to isolate and identify the decision and the link to its implications. Fourth, entrepreneurship is a fundamental driver of economic performance and innovation, albeit an uncertain and risky endeavor; interventions that increase the successful launch of new ventures

are therefore particularly relevant.

We develop a dynamic model of decision-making under uncertainty, showing that causal theories raise economic performance. We then empirically test the predictions of our model to corroborate its logic. Our study analyzes data from eight pre-registered randomized control trials (RCTs) involving 1,556 early-stage startups across eight countries (China, Colombia, India, Italy, Netherlands, Spain, Tanzania, and United Kingdom), spanning a total of three years.[1] We extend prior research by comparing the performance of startups in a control group (*C*) that undergoes no training and two groups trained in two distinct decision-making approaches: (i) Evidence-Based (*E*) Approach – in which entrepreneurs rely on iterative experimentation without an explicit causal framework; (ii) Theory-and-Evidence-Based (*TE*) Approach – in which entrepreneurs learn to develop and test causal theories before conducting experiments. The interventions consisted of six half-day, in-person learning sessions totaling 24 hours of instruction (plus take-home exercises to boost compliance), delivered over a four-month period. This is a substantial treatment for such early-stage ventures, comparable to intensive entrepreneurship bootcamps from business schools. Throughout all phases of our study, we paid close attention to prevent or mitigate common sources of bias such as selection, self-selection, and contamination.

Our experimental design addresses potential confounds. The two treatment groups shared approximately 75% of their curriculum, with the only difference being the focus on theory-building for the *TE* condition. We also undertook careful steps to prevent contamination and avoid potential instructor-specific effects. We collected data through surveys and structured interviews conducted by trained research assistants at six points over the period of the study, measuring outcomes including revenue, employment, productivity, and business termination.

Our econometric analysis employs several complementary approaches. We mainly rely on difference-in-differences specifications with startup and time fixed effects to estimate intention-to-treat effects on performance. We supplement these analyses with a battery of alternative spec-

[1] These RCTs were registered on the American Economic Association's Registry for Randomized Controlled Trials with registration numbers AEARCTR-0009082, AEARCTR-0007560, and AEARCTR-0007594. They were registered separately as different sources of funding became available for the trials.

ifications and robustness tests, including adjustments for potential serial correlation in our panel data, attrition, and multiple-hypothesis testing. Furthermore, we perform survival analyses using Kaplan-Meier estimates and Cox proportional hazards models to study termination patterns. To get closer to our theoretical mechanisms, we also develop novel measures of theory-based causal reasoning and empirical testing, which we use in two-stage regressions to shed additional light on our results.

Our findings indicate that, conditional on survival, entrepreneurs trained in the *TE* approach significantly outperform those in the *E* and *C* groups in terms of revenue and productivity. Specifically, theory-based startups exhibit 1.28-1.7 times higher sales growth rates compared to control firms (equivalent to an average increase of about €17,000-€46,000), driven by an increase in productivity rather than employment growth. The substantial percentage magnitude of these effects is partly attributable to the early-stage nature of the ventures in our sample, as the monetary effect sizes show. Given the low baseline revenues, initial growth rates tend to be disproportionally high. In fact, this really underlies the fact that causal reasoning can considerably increase startups' early growth. Additionally, both treatment groups manifest higher termination rates over time, suggesting that structured decision-making approaches help entrepreneurs make better-informed termination decisions.

Overall, our study provides evidence that, similar to scientists, entrepreneurs should prioritize the development of causal theories alongside experimentation, rather than relying purely on data-driven iterations. More generally, it shows that theory-based decision-making is not just an academic exercise, but an important determinant of economic performance.

MODEL

2.1 Actions, experiments, theories

We model the implications of the interventions of our RCTs. Early-stage entrepreneurs, such as those in our study, decide whether to remain in business depending on their expected future returns. Let $K = \frac{\pi}{r}$ represent set-up costs $K > 0$, and the future gross profits π and returns r when

the venture stabilizes. Gross profits are not easy to predict because of the vagaries of accounting profits. Early-stage entrepreneurs typically make their predictions using future sales, which can be identified more objectively. At the same time, sales are often proportional to gross profits – for instance, under perfect competition and a Cobb-Douglas production function, or under monopolistic competition with a constant elasticity demand function and linear costs. They are reasonable models for the kind of entrepreneurs that we deal with.

If Y denotes the future sales when the venture stabilizes, then $r = \frac{Y}{K}$, where to streamline our notation we normalized r by dividing it by the proportionality factor that links gross profits to sales. We assume that r is log-normally distributed, consistent with the fact that sales cannot be negative. This also accounts for the fact that most early-stage entrepreneurs earn low returns, and a few of them earn high returns. Let $\ln \frac{Y}{K} \equiv y = \mu + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon)$ and μ is the prior expected log-return. For simplicity, we normalize the outside option's returns to 1, which implies that they pursue the venture if they expect to obtain log-returns $\mu > 0$, otherwise they earn the outside option. Our entrepreneurs then expect to earn returns

$$v(\mu) = max(0, \mu) \tag{1}$$

After they settle on a prediction μ, entrepreneurs take actions to maximize profits, typically the optimal choice of inputs. We leave these choices in the background. Entrepreneurs in the *C* group do not develop theories or run experiments to seek evidence about μ. They hold a prior μ_0 and based on this prior they choose actions to maximize profits. Entrepreneurs in the *E* group experiment to update the prior μ_0, and then choose actions. Entrepreneurs in the *TE* group develop causal theories to learn about μ_0. They experiment to update μ_0, like the entrepreneurs in the *E* group, and choose their actions. In all three cases, if $v(\mu) > 0$, they pursue the venture, otherwise they earn the outside option.

Entrepreneurs in the *TE* group instead start by formulating a *theory*, which is a structured causal model that maps antecedents and relevant contextual factors into a distribution of possi-

ble outcomes. It specifies (i) a state space of variables linked by a causal-logical structure and (ii) beliefs/parameters that restrict this space by assigning plausibility and stochastic predictions, including uncertainty from unmodeled influences. *TE* entrepreneurs therefore identify a set of causal antecedents $\mathbf{x}' = \{x_1, x_2, \ldots, x_m\}$ and determine their prior as $\mu_0 = \mathbf{x}'\theta$, where θ is a vector of weights of the causal antecedents defined by the theory. Thus, a theory identifies the relevant causal antecedents $\mathbf{x}$ of the problem, and defines, subjectively, the weights.[2] The $\mathbf{x}$ proxy, for instance, for whether a technology works, customer preferences are high or low, competition or regulation are tight or weak. The realizations of $\mathbf{x}$ could be uncertain, and *TE* entrepreneurs collect information to make predictions about them. Since this is not going to change the implications of our discussion, we assume for simplicity that *TE* entrepreneurs observe $\mathbf{x}$.

Entrepreneurs in the *TE* group then have different priors depending on their choice of causal antecedents and the observed realizations. Entrepreneurs in the *E* and *C* group ignore $\mathbf{x}$. Therefore, the different realizations of $\mathbf{x}$ do not affect their prior. We assume that the counterfactual *E* or *C* entrepreneurs set a prior μ_0 equal to the expected μ_0 across different realizations of $\mathbf{x}$, given θ – that is, they set prior $\overline{\mu}_0 = \mathbb{E}_{\mathbf{x}\in\mathbf{X}}\,\mu_0$, where $\mathbf{X}$ is the space of $\mathbf{x}$. For example, *TE* entrepreneurs observe that the technology works, in which case their prior is higher than *C* or *E* entrepreneurs, or the technology does not work, in which case it is lower, and similarly for other causal antecedents.[3]

2.2 Equilibrium

Entrepreneurs update their priors by running experiments that yield, sequentially, signals s_i in stages $i = 1, 2, \ldots$. For simplicity, we assume that signals are unbiased but noisy estimators of y, that is $s_i = y + \eta_i$, with $\eta_i \sim \mathcal{N}(0, \sigma_\eta)$. Let μ_0 be the prior of y before the first signal s_1 in stage 1.

[2]This mimics our entrepreneurial training. As discussed in Appendix A, we taught entrepreneurs to formulate theories that identify causal antecedents. They form a "story tree" that links causal antecedents to their effects. As also discussed in Appendix A, it is not easy to teach causal reasoning to individuals unfamiliar with it. Therefore, we did not belabor them with details about how to identify weights. We assume that, once they identified causal antecedents for their theories, they formed naturally their own subjective beliefs about weights. For example, we noted that they often set equal weights.

[3]The experiments of *TE* entrepreneurs may test the causal links θ instead of merely predicting Y given μ_0. For simplicity, we keep this point in the background because *de facto* these tests also yield updates of μ.

In stage i, the updated μ_i is

$$\mu_i = \mu_{i-1} + \alpha_{i-1}(s_i - \mu_{i-1}) \tag{2}$$

where $\alpha_{i-1} \equiv \dfrac{\sigma_{i-1}^2}{\sigma_{i-1}^2 + \sigma_\eta^2}$, $\sigma_i^{-2} = \sigma_{i-1}^{-2} + \sigma_\eta^{-2}$, and $\sigma_0 = \sigma_\varepsilon$. [4]

Entrepreneurs solve a dynamic program in which at the end of stage i they receive a signal s_i and choose whether to commit and obtain $v(\mu_i)$ or to run another experiment in stage $i+1$ to obtain a new signal s_{i+1}. The value at the beginning of each stage is

$$V_i(\mu_i) = max\left\{v(\mu_i),\ \beta\overline{V}_{i+1}(\mu_i) - c\right\} \tag{3}$$

where $\beta \in (0,1)$ is a discount factor and c is the cost of running an experiment to elicit s_{i+1}.[5] Let $\{\mathscr{F}_i\}_{i\geq 1}$ be the natural filtration generated by the signals, where $\mathscr{F}_i = \sigma(s_1, s_2, \ldots, s_i)$. The continuation value is equal to

$$\overline{V}_{i+1}(\mu_i) = \mathbb{E}\left(y \mid \mathscr{F}_i,\, s_{i+1} > s_{i+1}^*\right)\mathbb{P}\left(s_{i+1} > s^* \mid \mathscr{F}_i\right) \tag{4}$$

where s_{i+1}^* is the threshold level of the signal such that $\forall s_{i+1} > s_{i+1}^* \implies \mathbb{E}(y \mid \mathscr{F}_{i+1}) > 0$. The logic of this expression is that at the beginning of stage $i+1$ entrepreneurs expect to earn returns μ_{i+1}, which updates μ_i, if the new signal s_{i+1} is greater than the threshold, or 0 otherwise. In $V_i(\mu_i)$ they compare this continuation value that they obtain if they run the experiment, with $v(\mu_i)$, which is what they obtain if they do not run the experiment. The conditional expectation of y given $\mathscr{F}_{i+1}$ is (2) evaluated at $i+1$ rather than i, which implies $s_{i+1}^* = -\dfrac{(1-\alpha_i)\mu_i}{\alpha_i}$. The final conditional expectation (4) becomes

$$\overline{V}_{i+1} = \mu_i \Phi\left(\frac{\mu_i}{\lambda_i}\right) + \lambda_i \phi\left(\frac{\mu_i}{\lambda_i}\right) \tag{5}$$

where $\lambda_i \equiv \alpha_i\sqrt{\sigma_i^2 + \sigma_\eta^2}$.

[4] For simplicity we assume that entrepreneurs know the variances σ_ε^2 and σ_η^2

[5] As shown below, $\overline{V}_{i+1}$ is the expected rate of return before running a new experiment, which is expected profits relative to K. Thus, c is also the cost of the experiment relative to K

We solve this dynamic program in two steps. First, the lemma shows that the program has a finite time horizon that ends with the choice to abandon or commit to the venture, obtaining $v(\mu) = 0$ or μ. Second, the proposition provides the conditions under which they choose to run a new experiment in the next stage, or to terminate by picking the outside option or committing to the venture.

Lemma 2.1. $\exists\, i^* : i > i^* \rightarrow V_i(\mu_i) = v(\mu_i)$ & $\dfrac{\partial V_i(\mu_i)}{\partial i} \leq 0$

Proof. We prove this lemma in three steps. First, we show that the experiment at the beginning of any stage $i \geq 1$ is more valuable than any experiment in later stages, making $\beta\overline{V}_{i+1} - c < v(\mu_i)$ a sufficient condition to stop the process. At the beginning of stage i, $y = \mu_{i-1} + \varepsilon_i$. Since $s_i = y + \eta_i$, then $\mathbb{E}s_i = \mu_{i-1}$. From (2), this implies $\mathbb{E}\mu_i = \mu_{i-1}$. More generally, the process is a martingale with respect to the filtration $\mathscr{F}_i$, $\mathbb{E}(\mu_{i+k} \mid \mathscr{F}_i) = \mu_i$, $\forall k \geq 1$. Therefore, the expected outcomes of the current and future experiments is the same because, before running the experiment, they rely on the same information. Second, we show that $i \rightarrow \infty \implies \overline{V}_{i+1} \rightarrow \mu_i$. Recursive application of the formula for α_i implies $\alpha_i = \dfrac{\sigma_\varepsilon^2}{(i+1)\sigma_\varepsilon^2 + \sigma_\eta^2}$ and $\sigma_i^2 = \frac{\sigma_\varepsilon^2\sigma_\eta^2}{i\sigma_\varepsilon^2 + \sigma_\eta^2}$. It is not difficult to see that this implies that $i \rightarrow \infty \implies \lambda_i \rightarrow 0$. From (5), this implies $\overline{V}_{i+1}(\mu_i) \rightarrow \mu_i$. Finally, $s_i > s_i^* \implies \mu_i > 0$, which implies that entrepreneurs keep experimenting if $\beta\overline{V}_{i+1}(\mu_i) - c > \mu_i$. But if $\overline{V}_{i+1}(\mu_i) \rightarrow \mu_i$, for i sufficiently large, $\beta\overline{V}_{i+1}(\mu_i) - c < \mu_i$. Similarly, $s_i < s_i^* \implies \mu_i < 0$, , which implies that entrepreneurs keep experimenting if $\beta\overline{V}_{i+1}(\mu_i) - c > 0$. But if $\overline{V}_{i+1}(\mu_i) \rightarrow \mu_i$, for i sufficiently large, $\beta\overline{V}_{i+1}(\mu_i) - c < 0$ because $\overline{V}_{i+1}(\mu_i) \rightarrow \mu_i < 0$. □

Proposition 2.2. *Assuming that β is sufficiently large, or c is sufficiently small, $\exists\, (\mu_{i*}, \mu_i^*)$ with $\mu_{i*} < \mu_i^*$: (i) $\mu_i < \mu_{i*} \implies$ entrepreneurs terminate the venture and obtain $v(\mu_i) = 0$; (ii) $\mu_{i*} < \mu_i < \mu_i^* \implies$ entrepreneurs experiment to elicit a new signal s_{i+1}; (iii) $\mu_i > \mu_i^* \implies$ entrepreneurs commit to the venture and obtain $v(\mu_i) = \mu_i$.*

Proof. We prove this proposition in four steps. First, from (5), it is not difficult to see that $\dfrac{\partial\overline{V}_{i+1}}{\partial\mu_i} = \Phi\left(\dfrac{\mu_i}{\lambda_i}\right) > 0$. Second, when $\mu_i = 0$, $\overline{V}_{i+1}(\mu_i) = \lambda_i\phi(0) > 0$. For β sufficiently large or c sufficiently small, this implies that $v(\mu_i) = 0 < \beta\overline{V}_{i+1}(0) - c$. Third, $\mu_i \rightarrow -\infty$ implies that

$\overline{V}_{i+1}(\mu_i)$ becomes non-positive. In turn, this implies that for μ_i sufficiently small, $\beta\overline{V}_{i+1}(\mu_i) - c < 0$. Finally, $\Phi\left(\frac{\mu_i}{\lambda_i}\right) < 1$ implies that, when $\mu_i > 0$, $\beta\overline{V}_{i+1}(\mu_i) - c$ increases at a lower rate than $v(\mu_i) = \mu_i$. This ensures that for μ_i sufficiently large, $\beta\overline{V}_{i+1}(\mu_i) - c$ cuts $v(\mu_i)$ from below. The first three steps imply that $\exists \mu_{i*} : \mu_i < \mu_{i*} < 0 \implies \beta\overline{V}_{i+1}(\mu_i) - c < 0$. The fourth step implies that $\exists \mu_i^* : \mu_i > \mu_i^* > 0 \implies \beta\overline{V}_{i+1}(\mu_i) - c < \mu_i$. □

The intuition of the lemma is that, in a standard Bayesian Plausible Experiment (Moscarini and Smith, 2001; Frankel and Kamenica, 2019), as entrepreneurs run more experiments, the information content of the experiments is reduced. The updates become closer, and if experiments are costly or delay the outcome, at some point it is not worth running experiments anymore. The entrepreneurs of our RCTs are small entrepreneurs not in novel or frontier high-tech businesses. They explore in known state spaces, making this assumption plausible.

Our proposition then says that when μ_i is low, entrepreneurs commit to the outside option; when μ_i is high, they commit to the venture; in between, where the uncertainty is higher, they experiment. The new signals update μ_i upward or downward. If the updated $\mu_i + 1$ falls outside the interval (μ_{i*}, μ_i^*) they either commit or terminate, otherwise they keep experimenting.

Figure I provides a graphical representation. As the figure shows, $V_i(\mu_i)$ increases with μ_i and generates three zones: termination ($\mu_i < \mu_{i*}$); exploration ($\mu_{i*} < \mu_i < \mu_i^*$); and commitment ($\mu_i > \mu_i^*$). As the stage i increases, our lemma shows that the experimentation zone shrinks, making termination or commitment more likely. The parameters β and c affect the position of the curve. Higher β or lower c shift the expected value of experimentation upward, enlarging the experimentation zone. The dotted curve in the figure denotes the expected value of experimentation when β or c is such that there is no exploration zone. If β is sufficiently small or c is sufficiently large, entrepreneurs do not experiment.

2.3 Performance

The following theorem draws the main empirical prediction of our model.

Theorem 2.3. *Causal theories generate higher expected performance.*

Proof. We prove this theorem in three steps. First, consider an entrepreneur who identified causal antecedents $\mathbf{x}$ that generate heterogeneous priors $\mu_0 = \mathbf{x}'\theta$ depending on the realizations of the causal antecedents. The counterfactual is the same entrepreneur who did not develop a theory – that is, who did not identify the causal antecedents. We assume that, before observing the causal antecedents, the expected prior of the former entrepreneur is the same as the counterfactual entrepreneur, that is, $\overline{\mu}_0 = \mathbb{E}_{\mathbf{x}\in\mathbf{X}}\mu_0$. Since this is the same entrepreneur under the two conditions (developed a theory or not), at stage i they observed the same signals. The equality of expectations holds: $\forall\, i \geq 1$, $\overline{\mu}_i = \mathbb{E}_{\mathbf{x}\in\mathbf{X}}\mu_i$ because the recursive application of (2) and α_i imply $\mu_i = \mu_0 + \dfrac{i\sigma_\varepsilon^2}{i\sigma_\varepsilon^2 + \sigma_\eta^2}(\bar{s}_i - \mu_0)$, where $\bar{s}_i$ is the average value of the signals observed up to i. Second, from the proof of our proposition, it is easy to see that $\overline{V}_{i+1}(\mu_i)$, and therefore $\beta\overline{V}_{i+1} - c$ is convex in μ_i. Because $v(\mu_i)$ is linear in μ_i, $V_i(\mu_i)$ is also convex in μ_i. Third, this implies that $\mathbb{E}_{\mathbf{x}\in\mathbf{X}}V(\mu_i) \geq V(\mathbb{E}_{\mathbf{x}\in\mathbf{X}}\mu_i)$. Therefore, before observing the causal antecedents, entrepreneurs who develop a causal theory of their performance expect a higher value than the counterfactual entrepreneurs who do not develop a theory. □

As the proof of the theorem shows, *TE* entrepreneurs perform better because they generate heterogeneous priors produced by the variation in the causal antecedents, and not just in the updates of μ_i from signals. The convexity of $V(\mu_i)$ implies that the upside produced by favorable realizations of causal antecedents carries a greater weight than the downside produced by unfavorable realizations. This convexity is produced by the convexity of the expected value of the experiment, which depends on the fact that the outside option generates a lower bound to the losses produced by negative outcomes from experimentation. Since this benefit is higher the greater the variability of the final outcome, theories, which generate this higher variability, raise expected performance. Simply put, by raising priors, favorable causal antecedents raise the upside from experimentation, while the lower priors generated by unfavorable causal antecedents are still sheltered by the outside option.

We also have the following result.

Result 2.4. Experimentation generates higher expected performance.

The proof of this result is straightforward. Without experimentation, entrepreneurs in the *C* group expect $V(\mu_i) = v(\mu_i)$ because they ignore the curve $\beta\overline{V}_{i+1}(\mu_i) - c$. As shown in our proofs and in Figure I, this raises $V(\mu_i)$ compared to entrepreneurs who run experiments in the range (μ_{i*}, μ_i^*). Thus, experimentation raises expected performance.

Finally, we make the following conjecture.

Conjecture 2.5. *Causal theories and experimentation generate higher termination rates.*

We cannot make unambiguous conclusions about termination. We would need to make assumptions about μ_0 and the parameters β and c that determine μ_{i*}. Since we have no basis to make these assumptions, we prefer to state our prediction about termination as a conjecture.

The experimentation curve observed by *TE* and *E* entrepreneurs generates two contrasting effects. First, at any stage i, they are less likely to terminate than *C* entrepreneurs because there is a range $\mu_{i*} < \mu_i < 0$ in which they experiment while *C* entrepreneurs terminate. Second, we can make the fair assumption that *TE* and *E* entrepreneurs update their probabilities more frequently because we induced them to experiment while entrepreneurs in the *C* group acquire information only when they receive exogenous signals during their operations. Our lemma implies that the former entrepreneurs are more likely to terminate or commit earlier. This effect dominates if, in a given time horizon, the frequency of experimentation is sufficiently higher.

On top of experimentation, the effect of causal theories is also ambiguous. Causal theories generate a higher oscillation band of priors, and therefore of updates. The lower prior is more likely to generate termination. However, this effect is offset by the higher prior. The net effect is unclear, leading us to study our conjecture empirically.

DATA SOURCES AND EXPERIMENTAL SETTING

The data for this paper were collected through eight field RCTs involving a total of 1,556 early-stage startups[6] based in eight countries (in chronological order: Tanzania, China, The Netherlands, Italy, Colombia, United Kingdom, Spain, India; see Table I for more details). These countries comprise both advanced and developing economies, with significantly different cultures, norms, resources, regulations, and entrepreneurial ecosystems, resulting in a rich, generalizable, and unique set of data.

Each of the eight trials involved a data collection period of approximately 15-17 months, comprising pre-treatment, during-treatment and post-treatment measurements (Figure II). Minor differences in timings across RCTs are due to location-specific reasons and festivities. The only significant difference is represented by the last RCT (India), where data was collected for a period just shy of ten months due to logistical and institutional reasons. Overall, our data collection efforts span a three-year period from April 2021 to April 2024 and four continents. Each of the eight RCTs was conducted following the same structure and protocols, also involving partnerships with local institutions to facilitate the successful realization of the trials. Local partners and teams were instrumental, among other things, in tailoring and promoting the marketing campaigns, providing access to facilities, and recruiting instructors and research assistants (RAs henceforth), following standardized procedures from the research team.

3.1 Recruitment of participants

Participants were separately recruited at each RCT location through marketing campaigns that combined social media advertisements, word-of-mouth referrals within entrepreneurial networks, and endorsements from institutions such as universities and business associations. The calls for applications were disseminated via dedicated websites (sharing a common interface and content), targeting early-stage entrepreneurs. Application eligibility was not restricted based on demographic

[6]Due to local specificities, the first two RCTs also allowed more established small businesses to apply to the program with an innovative project that they wanted to start developing.

characteristics or industry affiliation. To minimize selection bias, the program was advertised as a general entrepreneurship pre-acceleration initiative, avoiding emphasis on specific topics that might disproportionately attract certain applicants. The websites and marketing materials were purposely general and did not contain any details regarding the experimental design of our study (e.g., existence of different treatment conditions, content of the treatments, etc.). In addition, participation was offered free of charge to prevent self-selection based on financial constraints. Interested entrepreneurs were required to complete an online application, submit a baseline questionnaire, provide a written description of their innovative business idea, and participate in an interview with one of our RAs. All startups that successfully completed these steps were admitted, ensuring that the selection was not influenced by subjective criteria. Admitted participants underwent up to five more rounds of data collection through interviews and surveys – approximately once every two months – to build a panel data set.

3.2 Randomization and experimental design

For each RCT, we used simple randomization to allocate participants into the three experimental conditions: *Theory-and-Evidence-based (TE)* treatment group, *Evidence-based (E)* treatment group and a *Control (C)* group (as in Agarwal et al., 2025). Our randomization script was designed such that the two treated groups would have more startups (i.e., a larger allocation ratio to treatment conditions compared to control), in order to reach a higher statistical power when estimating differences between theory and experimentation approaches, which represents the key goal and contribution of our study.

To prevent contamination between conditions, we asked admitted applicants to report acquaintances with other participants and then assigned those who did to the same treatment group. Only a handful of participants required reassignment, posing minimal risk to the integrity of the randomization process. Randomization was performed after participants completed the baseline data collection, to avoid possible treatment-specific self-selection bias (Floyd and List, 2016) [7].

[7] As described above, participants were not even aware of the existence of an experimental design or different

To enhance learning and intervention effectiveness, we further randomized the two treatment sub-groups into smaller classes of approximately 20-25 startups each, and randomly assigned each class to one of our 27 experienced instructors (i.e., three or four per RCT). Each instructor was carefully selected and underwent approximately 20 hours of training – standardized across RCTs – to ensure consistent delivery of the treatment contents in alignment with the experimental design. To maintain uniform quality and prevent instructor-specific biases in our inference, every instructor was assigned to both treated sub-groups. Additionally, to mitigate potential contamination, the two treatment sub-groups were scheduled for training in different consecutive days or at different times of the day (i.e., morning vs. afternoon), preventing any direct interaction. Communication channels, including emails, social media groups, and chat platforms, were also strictly segregated between all groups. These measures were implemented to prevent participants' awareness of different treatment conditions and to strengthen the internal validity of the study.

The *E* treatment sub-group received training emphasizing experimentation as a decision-making tool, while the *TE* sub-group was also taught how to develop causal theories of value alongside experimentation. The two sub-groups shared approximately 75% of the curriculum as measured by overlapping slides and core modules, with the remaining 25% of the formal materials tailored to either causal theory-building (*TE* condition) or additional experimentation practice (*E* condition). In the *TE* condition, however, causal theory-building was not confined to this distinct subset of materials; it provided the organizing framework for the full training sequence, such that instructors systematically connected the shared tools, examples, and exercises to participants' evolving causal theory throughout each session and related take-home work. Accordingly, the effective contrast between *TE* and *E* reflects a sustained difference in how the shared content was structured and interpreted, rather than a small isolated block of instructional time.

All participants in the treatment groups attended, free of charge, a four-month training program, while those in the control group received only informational content, such as news, videos, and webinars, without any formal training[8]. The interventions consisted of six half-day, in-person

conditions.

[8]These materials were also given to the treated entrepreneurs, to prevent possible confounding effects.

learning sessions, totaling 24 hours of instruction – plus additional take-home exercises to boost treatment compliance and effectiveness. Training sessions were delivered on average every other week, leaving participants enough time to absorb and apply the treatment materials to their ventures. Figure III provides a brief overview of the key contents of the six four-hour training sessions of each treated sub-group, while Appendix A provides important in-depth information on the treatments and their materials.

The training was designed to help entrepreneurs assess the viability of their business ideas. The first three sessions focused on strategy validation, introducing participants to widely used frameworks such as the business model canvas, customer personas and customer journey mapping, as well as data collection through interviews and questionnaires. Sessions four and five addressed solution validation, teaching participants how to develop a minimum viable product (MVP) and test it with customers. The final session covered pricing strategies, summarized the key course takeaways, and concluded with a pitching session where participants rehearsed investor presentations. The training was highly interactive, and incorporated real-world examples and a case study to bridge theoretical concepts with practical applications. The only difference between the two treated sub-groups regarded how to apply causal reasoning to develop theories of value: in the *TE* group, participants were taught to develop theories and to use them as a guiding principle for all aspects of decision-making, whereas in the *E* group, theories were absent. Importantly, causal theories acted as the fulcrum of the *TE* treatment, to which all other training contents were linked, including ones that would be otherwise identical to the *E* treatment (e.g., testing, data collection, etc.). To further increase the effectiveness of the treatments, participants were given assignments after each class designed to help apply the training contents to their specific ventures. Appendix A provides further details on our experimental treatments.

Following the four-month training program, participants were invited to attend several additional events, offered free of charge both in person and via live streaming. These events featured guest speakers covering topics complementary to the core training, such as social media marketing, intellectual property protection, digital content management, and accounting. Designed as an in-

centive to sustain engagement with both the program and data collection efforts, these events were made available to all participants, regardless of treatment assignment, with appropriate safeguards in place to prevent contamination across groups.

DATA AND KEY VARIABLES

Our study sample comprises 1,556 nascent and early-stage entrepreneurial ventures based in eight countries. At baseline, the median startup does not generate any meaningful revenue, employs two people (usually the co-founders), has a for-profit orientation, and has been operational for seven months. Around 74% of the sample sell services in the broad sense or a mix between products and services. The startups span all main industries and sectors, with the three most represented being agriculture and animal husbandry (12.34% of the sample), tourism (9.70%), and social media (8.03%). The median founder is 32 years of age, male, possesses a tertiary education degree, has seven years of prior work experience, and works part-time on the startup. Our sample is broadly representative of the larger entrepreneurial population, providing us with an ideal testbed to measure the real-world impacts of theory-based decision-making on the performance of new ventures.

We collected data for approximately 15 months for each RCT, forming a panel dataset with up to six data points per startup, covering pre-treatment (i.e., baseline), during-treatment, and post-treatment periods. Each data collection point involved systematically asking active participants to complete a survey questionnaire and conduct a structured interview with one of our purposely trained (and randomly assigned) RAs. RAs – undergraduate or graduate students selected from local partner institutions based on academic performance, knowledge of entrepreneurship, and analytical and communication skills – were recruited, tested, and trained to ensure consistency in data collection. Interview coding followed a standardized protocol, using a predefined code-book to assess the extent to which startups developed causal theories and applied experimentation in their activities (analogous to the process described by Bloom and Van Reenen, 2010 for coding management practices). In addition to conducting interviews and coding interview responses, other

RA responsibilities included assisting participants and the research team throughout the RCTs, supporting instructors during training sessions, and ensuring the completion of the survey.

The key outcome variables used in our study are the following: (i) *Termination*, which is a dummy equal to 1 if the startup terminated its activities and was no longer operating within the data collection period, and 0 otherwise[9]. Notably, attritors who simply stop participating in our data collection without mentioning having ceased their activities (e.g., they stop providing data due to personal reasons, lack of time, etc.) are not counted as having terminated their startup. Albeit likely to reduce our effect sizes, we opted for this more conservative assumption; (ii) *Cumulative Sales (Log €)*, which reports the cumulative amount of revenue in logged Euro earned by each startup at any data collection point. This approach was chosen to smooth out short-term fluctuations and seasonal effects that might otherwise add noise to the analyses, as well as to measure differences in growth rates, which are of primary interest for early-stage ventures. This is especially relevant when studying early-stage startups, whose sales can exhibit significant volatility in any single two-month data collection period. To mitigate potential concerns about serial correlation of the measure, we employ difference-in-differences estimators across our performance analyses and a battery of robustness checks; (iii) *Employees*, which measures the number of founders, full-time, and part-time employees working for the startup.

We also collected a number of other variables to verify that startups in our three experimental arms are balanced in terms of characteristics that might be potentially relevant. These variables include both firm-level measures – age of the startup in months, type of product or service sold, for-profit orientation, development stage, etc. – and founder-level ones – age, gender, education, work experience, weekly hours dedicated to the startup, etc. Crucially, we also collected data on two variables related to our theoretical mechanisms: (i) *Theory Score*, which captures the extent to which each startup's theory is clear, causal, well-grounded, and articulated; and (ii) *Empirics Score*, which captures the extent to which each startup collected data, ran tests and evaluated such evidence. Appendix B provides detailed information on the definition, components, measurement,

[9]This act is often also referred to as *"exit"* in the entrepreneurship literature. We also employ a panel measure of termination suitable for survival analysis models, as well as a variable capturing the time to termination.

and data collection of these variables.

Table II lists and briefly describes the key variables employed in the study. It also describes the data source that was used to collect them (i.e., surveys and/or interviews). Table III then provides the overall summary statistics for the aggregated sample at baseline (i.e., pre-intervention).

To verify that our separate random assignment procedures by RCT yielded an overall balanced sample when aggregated, we perform tests of balance. Table IV presents the means and standard errors (in parentheses) of our baseline covariates, split by treatment group. The last three columns report the p-values of the pairwise t-tests for each variable and treatment pair. Out of 51 pairwise tests (17 variables, 3 treatment pairs), only one of them returned a statistically significant difference at the typical 95% confidence level: *E* startups are on average 1.66 months older than control ones at baseline ($p = 0.03$). Such minor imbalance is to be expected for such a large number of pairwise tests, and the magnitude of the difference is not concerning[10]. We also perform F-tests for joint significance across all balance variables for each test pair, which result in low and not significant F-statistics (reported at the bottom of Table IV), indicating that our sample remains well balanced even when aggregating the eight RCTs. Tests reported in Table IV include standard errors clustered at the startup level, and fixed effects for each RCT because the randomization was performed within each location separately.

As is common with most RCTs and despite applying best practices to prevent it (Gerber and Green, 2012; Ghanem, Hirshleifer, and Ortiz-Becerra, 2023), our sample was affected by attrition. In particular, not all startups attended all six sessions of our training program or responded to all data collection surveys and interviews. Conditional on survival, our average response rate during the five post-baseline data collection periods was 80%, with a minimum rate of 76% in the first period and a maximum of 87% in the last. These response rates are also impacted by entrepreneurs who *terminate* their activities, one of our study's outcome variables, and therefore stop participating in our data collection. Net of these effects, the average attrition rate in the five post-baseline data collection periods (conditional on survival) was 13%, ranging from 19%

[10]Most of our regression specifications also employ startup-level fixed effects, which account for individual differences such as startup age at baseline.

in the first period to 9% in the fourth. Startups that chose to withdraw from the program and declined further interviews cited reasons such as time constraints, personal issues, or reluctance to continue sharing data. The initial peak of attritors is likely due to the free-of-charge nature of our study, which was aimed at minimizing self-selection biases based on income but, in turn, generated higher attrition rates because of the lack of penalties for non-motivated participants. Appendix C extensively discusses attrition, inspects for selective attrition, and presents robustness tests for our analyses, including inverse probability weights and different imputation techniques for missing data.

MAIN RESULTS

In this section, we set out to test our empirical predictions, leveraging our full dataset on 1,556 early-stage startups. Following our theoretical model, we start by estimating intention-to-treat (ITT) causal effects on key performance outcomes, in terms of revenue, employment, and productivity. We then examine ITT effects on termination, employing a variety of specifications. Finally, we use measures of theory and experimentation quality to explore our theoretical mechanisms.

5.1 Performance

Our theorem postulates that causal theories lead theory-based startups to perform better than the other groups, while experiments should generate higher performance for evidence-based startups compared to control ones. We therefore empirically test these predictions employing a difference-in-differences approach, estimating the following panel OLS regressions with startup and time fixed effects:

$$Y_{ijt} = \alpha + DD_{ijt} + S_i + T_t + \varepsilon_{ijt} \tag{6}$$

where Y_{ijt} is a vector of performance variables for startup i – assigned to treatment j – at time t, DD_{ijt} is the classic difference-in-differences estimator that indicates whether the treatment j (i.e., C, E or TE) has affected startup i at time t, S_i and T_t are fixed effects for startup and time, respectively, and ε_{ijt} is an error term (clustered at the startup level). The startup-level fixed effects

absorb all time-invariant individual characteristics, notably also including RCT-specific effects.

5.1.1 Revenue

We start by estimating the ITT effect of our treatments on startups' revenue, a key performance parameter for early-stage ventures. The main dependent variable in Table V is *Cumulative Sales (Log €)*, which reports the cumulative amount of revenue in logged Euro earned by each startup at any given data collection point. Sales values for RCT countries that do not adopt the Euro as a currency were converted to Euro using the appropriate average exchange rate for each local RCT currency and data collection period, taken from the European Central Bank's official website.

Column (1) in Table V presents results using the full panel of available data (5,270 startup-period observations). Thanks to the log transformation, the difference-in-differences coefficients are conveniently interpreted in terms of log ratios between sales growth rates, which are of primary interest for startups. Results in Column (1) show that *TE* startups' sales grow 1.7 times more than *C* startups $(exp(\beta) = 1.70, p = 0.001)$. The *E* treatment's coefficient indicates that revenue for startups in this group grow 1.06 times more than *C* ones, although this is not statistically significant $(exp(\beta) = 1.06, p = 0.694)$. As previously mentioned, the substantial magnitude of the effect observed in the *TE* experimental condition is partly attributable to low (or zero) initial sales of the early-stage startups in our sample, which makes percentage growth rates potentially higher compared to mature firms. In monetary terms, this corresponds to an average revenue increase of around €17,500 for *TE* ventures compared to *C* ones during our data collection period. For comparison, a related experimental study by Bloom et al., 2013 reported that a treatment in management practices led to a 17% increase in productivity, equivalent to over $300,000. However, their sample consisted of "large, multiplant Indian textile firms" (p. 3) with revenues in the millions of dollars.

To account for potential underestimation of standard errors stemming from serial correlation in our cumulative sales measure over time, Column (2) estimates the same regression as Column (1), this time using only a "pre-" (i.e., baseline) and a "post-treatment" period (i.e., the

last available data collection period). To avoid making inferences only on survivors in this specification, we prudently used the last available sales number for all startups, including those that terminate during the study period. This results in 3,112 observations in this two-period panel (1,556 startups, 'pre' and 'post'). Removing the time-series dimension sacrifices a significant amount of statistical power, but more accurately estimates standard errors (Bertrand, Duflo, and Mullainathan, 2004). Coefficients in Column (2) confirm that theory-based startups' sales grow significantly more than those of control ones ($exp(\beta) = 1.78, p = 0.000$), while the effect for evidence-based startups is once again positive but not statistically significant at conventional levels ($exp(\beta) = 1.24, p = 0.158$). The close alignment between the results of Columns (1) and (2) confirms that serial correlation in the dependent variable is not a significant concern in our study, given that our panel includes only a brief time series (i.e., six data points over approximately 15-17 months).

Given the skewed distribution of our revenue data, with 69% of startups having zero baseline sales, Column (3) in Table V then employs as dependent variable a binary indicator of whether a startup generates positive revenue at every data collection period. Following the same difference-in-differences model as Equation 6, Column (3) estimates that the *TE* treatment causes an increase of four percentage points (roughly 29%) in the probability of generating revenue ($\beta = 0.04, p = 0.030$), compared to control startups. On the other hand, the performance effect of the *E* treatment is practically null ($\beta = -0.00, p = 0.948$). Results from Column (3) therefore reinforce the interpretation that the outsize magnitude of the effects shown in Columns (1) and (2) are partly due to a significant proportion of observations starting from zero baseline revenue, with the *TE* treatment successfully bringing them to reach product-market fit. Given this premise, we are interested in assessing a more conservative estimate of the *TE* treatment's effect. To do so, Column (4) estimates the same model as Column (1), but only on the subsample of entrepreneurs who already had positive sales at baseline, significantly reducing skewness in the data[11]. Coefficients from Column (4) show that the *TE* startups' sales grew 1.28 times more than control

[11] After excluding the zeros, the *Cumulative Sales (Log €)* dependent variable closely resembles a standard normal distribution.

ones ($exp(\beta) = 1.28, p = 0.019$), giving us a lower-bound estimation of the *TE* treatment's performance effect. In monetary terms, this corresponds to an average differential revenue increase of around €46,000 during our approximately one-year observation period. However, the importance of starting to generate revenue should not be ignored. It constitutes an important milestone in entrepreneurial development and is often interpreted as a proxy for attaining product-market fit, securing customer traction, and successfully transitioning from "zero to one" (Thiel and Masters, 2014). Once again, even in Column (4) the *E* treatment does not appear to have any significant effect ($exp(\beta) = 1.05, p = 0.547$) on performance.

When comparing the two treatment coefficients, Wald tests of equality reject the null hypothesis that they are equal ($p = 0.002$, $p = 0.005$, $p = 0.015$, and $p = 0.042$ for the specifications of Columns (1), (2), (3), and (4), respectively). This confirms that the revenue growth rates of the theory-based group are larger than those of both the control and the evidence-based groups' ones. Given that the primary interest of this paper is studying the differential performance effects of formulating causal theories, we perform further analyses contrasting the *TE* and *E* groups. By directly comparing the two treated groups – which attended the same number of training sessions, taught by the same instructors, in the same locations, at virtually the same time – we also seek to isolate the effect of causal theories and eliminate concerns that our results in Table V might be due to so-called 'Hawthorne effect'.

Figure IV presents the cumulative distribution functions (CDFs henceforth) of the *Cumulative Sales (Log €)* outcome variable for *TE* and *E* startups. Panel A plots the CDFs for the baseline and during-treatment periods, while Panel B plots the CDFs for the post-treatment periods, consistent with our difference-in-differences regressions in Table V. Panel A of Figure IV shows that the revenue distribution of the two groups is virtually identical at baseline and during treatment (two-sample Kolmogorov–Smirnov test results: $D = 0.01, p = 0.999$), confirming the effectiveness of our randomization and the lack of differences while the startups were being treated. Panel B then shows that, conditional on survival, the post-treatment CDF of the *TE* group exhibits first-order stochastic dominance over that of the *E* group. This indicates that the *TE* treatment shifted the

sales distribution upward, compared to the E treatment (two-sample Kolmogorov–Smirnov test results: $D = 0.09, p = 0.005$).

To further investigate the time dynamics of the differential sales growth rates between the two treatment groups, we estimate the following regression on the subset of startups assigned to the TE and E groups:

$$log(CumulativeSales)_{ijt} = \alpha + TE_{ij} \times Months_t + S_i + T_t + \varepsilon_{ijt} \tag{7}$$

where we interact the treatment with the time variable (i.e., months since the start of the interventions). S_i and T_t are still fixed effects for startup and time, respectively, and ε_{ijt} is an error term (clustered at the startup level). This way, we present a more granular analysis of the treatment effects over time. Figure V plots the coefficients of this regression, where the blue curve represents the TE treatment and the counterfactual (i.e., the thick red line in $y = 0$) is the E treatment. This graph confirms that there are no significant differences between the two groups while they are being treated, further ruling out the possibility of Hawthorne effects or other biases introduced by our interventions. Moreover, Figure V shows that, around seven months after the start of the interventions (just three months after the end of the training), the revenue growth rate of theory-based startups starts to be significantly higher than that of evidence-based ones.

These results are consistent with our theoretical model's predictions, since theory-treated startups concentrate their efforts on building causal models to refine and update their priors during the intervention period, and they then commit at a higher expected performance level. These results therefore support our theorem, confirming that theory-based reasoning generates higher entrepreneurial performance. On the other hand, evidence-based entrepreneurs run tests and experiments to collect signals and concentrate their priors; however by not having a causal theory, they might misinterpret the relationships in the data. Our results in Table V show that experimentation might be slightly beneficial for evidence-based startups compared to control ones, in line with the result of our theorem. This positive effect is however not statistically significant, poten-

tially indicating that the position of the experimentation curve in Figure I is too low (i.e., low β and/or high c) to generate a substantial performance improvement without the guidance of a causal theory. The popularity of A/B tests and other experimentation techniques in the entrepreneurship practice and online resources might also lead some control entrepreneurs to get familiar with those procedures, rendering the counterfactual more challenging.

5.1.2 Employment and productivity

Next, to deepen our understanding of performance effects, we ascertain whether our causal theory treatment increased startups' output (i.e., cumulative sales) by increasing productivity. Table VI presents the results of difference-in-differences ITT regressions estimating employment and productivity outcomes, following the same specifications as in Equation 6.

Column (1) of Table VI shows that both *TE* and *E* treatments do not increase the number of startup employees compared to control firms ($\beta_{TE} = 0.20, p = 0.404; \beta_{EB} = 0.21, p = 0.418$). Further analysis of turnover reveals that most startups neither added nor removed employees during the study period, consistent with their status as nascent entrepreneurial ventures primarily composed of founding team members. This implies that the positive effect of the *TE* treatment could be due to an increase in productivity. Column (2) indeed shows that the productivity of *TE* startups (measured in Euro of revenue per employee) increased 1.86 times more than that of *C* startups ($exp(\beta) = 1.86, p = 0.001$). The ITT effect of the *E* treatment on startups' productivity is instead positive but not statistically significant ($exp(\beta) = 1.12, p = 0.482$). Wald tests of equality also confirm that the difference between the two treatment coefficients is statistically significant ($p = 0.004$). To increase robustness in our productivity analyses, we calculate a residual proxy measure of total factor productivity (TFP) by regressing the logged cumulative sales on the log number of employees, with time and RCT dummies. Given that most startups in our sample have very little to no capital, we believe that this is a reasonable assumption. We then predict the residuals of this regression and use them as dependent variable for our difference-in-differences regression (as in Equation 6, with bootstrapped standard errors to account for the

usage of residuals). Column (3) reports results of this proxy measure of TFP, indicating that *TE* treatment caused a 1.69 times higher increase in productivity compared to the control group ($exp(\beta) = 1.69, p = 0.001$). The *E* treatment's effect is again positive but not statistically significant ($exp(\beta) = 1.14, p = 0.380$). As above, the Wald tests reject the null hypothesis that the coefficients of the two treatments are equal ($p = 0.011$). Overall, these results further corroborate our theoretical model, indicating that theory-based reasoning generates higher performance by increasing the entrepreneurs' productivity. Experimentation alone, as indicated by comparing *E* and *C* entrepreneurs, has on average a positive but non-significant impact on productivity. Moreover, while productivity increases often lead to higher employment, this is a medium- to long-term effect, which we do not observe in our approximately 1.5-year data collection window.

To verify robustness and increase confidence in the results presented in these sections, Appendix C includes in-depth analyses and countermeasures for attrition, Appendix D shows that results hold even after applying multiple hypothesis testing adjustments and Fisherian randomization inference, and Appendix E explores heterogeneity in our treatment effects employing state-of-the-art machine learning techniques.

5.2 Termination

After testing our key theorem and its result, we empirically assess our conjecture and examine whether our interventions also make entrepreneurs more likely to terminate their activities. Given that our theoretical predictions on this matter are ambiguous, rigorous empirical evidence is especially valuable.

Figure VI shows a Kaplan-Meier survival graph, plotting the survival probability over time of all startups in the three experimental groups. Upon initial examination, the first noticeable difference across the three curves is recorded in the first post-baseline data collection period (i.e., 2 months after the start of the interventions), when the control group exhibits a substantially larger rate of termination compared to treated startups. When interviewed by our RAs, most of these entrepreneurs mentioned having quit their business because it was not sustainable without exter-

nal financing – a well-known finding in the literature (Holtz-Eakin, Joulfaian, and Rosen, 1994; Cagetti and De Nardi, 2006). After this initial phase, the curves in Figure VI gradually become closer and overlap, until around $t = 10$ months when the two treated groups start exhibiting greater termination rates. Log-rank tests of the overall difference in termination between groups show weak support when comparing the *C* group with the *E* group ($\chi^2 = 2.79, p = 0.095$) or with the *TE* group ($\chi^2 = 2.60, p = 0.107$).

Overall, the patterns highlighted in Figure VI are consistent with the reasoning behind our conjecture. At earlier stages, *C* entrepreneurs terminate more because there is a range $p_* < p < \tilde{p}$ (recall Figure I) in which they terminate while *TE* and *E* ones experiment. As the time horizon widens, treated startups – because of experimentation – receive more signals, engage in more frequent terminate-experiment-commit decisions compared to *C* entrepreneurs, and eventually are more likely to terminate. Our empirical evidence suggests that, after approximately $t > 10$ months, the experimentation effect dominates, so that treated entrepreneurs terminate more than control ones, consistent with our conjecture. Additionally, a comparison between the *TE* and *E* curves indicates that they are not statistically different from each other ($\chi^2 = 0.00, p = 0.964$). As discussed in our conjecture, this result suggests overall symmetry in the oscillation band of priors for *TE* entrepreneurs vis-à-vis *E* ones, with the effect of the lower priors (greater termination) being offset by the higher priors (less termination).

To supplement the Kaplan-Meier analyses in Figure VI, we perform additional regressions to test the ITT effect of our treatments on termination rates. We start by estimating a cross-sectional linear probability model of termination, whose results are reported in Column (1) of Table VII. Both treatments exhibit a five percentage points increase in the probability of termination, compared to the control group ($p \approx 0.03$ for both coefficients). Probit average marginal effects in Column (2) are virtually identical to the OLS ones reported in the first column, although with stronger statistical significance ($p \approx 0.01$ for both coefficients). To account for the right-censoring of our survival data, we then estimate the time to termination through a Tobit model. Estimates in Column (4) show that treated startups cease their activities around 2.8 months earlier than control

firms ($p \approx 0.04$ for both coefficients). These results are likely driven by the higher post-treatment termination rates shown in Figure VI that outweigh the initial terminations by the control group. Overall, these results indicate that, for sufficiently ample time horizons, *TE* and *E* entrepreneurs are similarly likely to terminate their activities and more likely than *C* ones, consistent with our conjecture. Terminations are key to a more efficient allocation of resources in the economy, as entrepreneurs often exit their ventures to pursue better or more remunerative business opportunities (Global Entrepreneurship Monitor, 2025).

We then turn to our full panel dataset and estimate a Cox proportional hazards model, which also accounts for the time period in which each startup terminates. The sample size for this model is the same as that for the Kaplan-Meier plots ($N = 6,867$), since both analyses follow a single-failure-per-subject approach. This means that each startup can only experience the event once (i.e., *Termination*), after which they are removed from the risk set. Estimates from this Cox model in Column (3) of Table VII suggest that both treatments increase the hazard of terminating by around 31-32%, although this time the effects are not statistically significant ($p \approx 0.13$ for both coefficients). Notably, the lack of observable differences in termination rates between startups in the *TE* and in the *E* treatments – apparent from Figure VI and all columns of Table VII – alleviates potential concerns when conditioning on survival for the analyses in our previous sections.

5.3 Theory quality, experimentation, and performance

A central role in our theoretical model is played by entrepreneurs' causal theories, which enable startups to identify optimal choices about causal antecedents that other entrepreneurs do not see. Given that the *TE* and *E* treatments differed only by the formulation of causal theories and were otherwise identical, our empirical findings so far represent evidence compatible with the fact that entrepreneurs' causal theories drive positive performance effects.

In order to get closer to our theorized potential mechanisms, we start by checking whether – following our experimental design – the *TE* treatment indeed differed from the *E* treatment only in terms of causal theories and not experimentation. We therefore estimate the following two

regressions on the subsets of startups assigned to the two treated groups:

$$Theory\ Score_{ijt} = \alpha + TE_{ij} \times InterviewRound_t + S_i + I_t + \varepsilon_{ijt} \tag{8}$$

$$Empirics\ Score_{ijt} = \alpha + TE_{ij} \times InterviewRound_t + S_i + I_t + \varepsilon_{ijt} \tag{9}$$

where we interact the *TE* treatment indicator with the data collection interview round (because the two dependent variables are collected through periodic interviews). S_i and I_t are the fixed effects for startup and interview round, respectively, and ε_{ijt} is an error term (clustered at the startup level). As described in Table II, *Theory Score* captures the extent to which entrepreneurs formulate clear, well-founded, and causal theories. *Empirics Score* instead quantifies the degree to which each startup gathered relevant evidence (e.g., through experiments, interviews, surveys, or MVP testing), analyzed the collected data, and made decisions informed by the resulting insights. These indexes are computed as the cumulative evaluation assigned at each data collection point by our RAs after standardized interviews with startups, and measure two crucial aspects of our theoretical mechanisms. Appendix B contains detailed information on how these variables are defined and on how the relevant data were collected. Figure VII plots the coefficients of these regressions, where the blue curves represent the *TE* treatment and the counterfactual (i.e., the thick red line in $y = 0$) is the *E* treatment. Following our experimental design, Panel A shows that the *TE* treatment causes a significant increase in the causality, clarity, and quality of startups' theories, compared to the *E* treatment. Panel B then shows that there are no measurable differences between the treated groups with respect to the quality and extent of the tests and evidence gathering. These results reinforce our confidence in the validity of our treatments and therefore are an additional step toward attributing the previous evidence of differential performance effects of the *TE* treatment to the formulation of causal theories.

To further investigate the relationship of theories and experiments with performance, we then follow Bloom et al., 2013 in exploiting experimental variation in treatments to examine whether our theory and empirics indexes are empirically linked to performance outcomes. While

we cannot support the exclusion restriction behind using these instruments, the 2SLS approach presented here can be interpreted as a step to verify whether our 'theory' and 'empirics' indexes are indeed correlated with performance. Appendix B reports first stage difference-in-differences regression results on the said scores for treatments' validity. Results in Column (1) of Table VIII suggest that a one-unit increase in *Theory Score* leads to a 134% increase of startup cumulative sales ($exp(\beta) = 2.34, p = 0.000$). This is both statistically and economically significant, although this large effect is once again likely skewed by the majority of startups not generating any baseline sales. Estimates in Column (1) for the *Empirics Score* index instead show a negative coefficient when keeping the *Theory Score* constant, although it is only mildly statistically significant ($\beta = -0.76, p = 0.082$). One possible explanation could be that startups that conduct more tests are delaying the generation of revenues during our study's time frame to gather more information in hope of higher future returns.

We then also estimate the same 2SLS model for employment and productivity outcomes. Column (2) in Table VIII shows that neither the 'theory' nor the 'empirics' score indexes are significantly related to the total number of people working for the startups, in line with results from Table VI. Estimates from Column (3) in Table VIII then show that better theories are correlated with a significant increase in productivity, as proxied by the log of the ratio between cumulative sales and the total number of employees. A one-unit increase in *Theory Score* is associated with a 129% increase in the amount of sales per start-up employee ($exp(\beta) = 2.29, p = 0.043$). Similarly to Column (1), also here the estimated *Emprics Score* coefficient is negative, although not statistically significant ($\beta = -0.71, p = 0.111$).

These results reinforce our findings that theory-based decision making is associated with significant performance benefits for entrepreneurs in our diverse sample. Moreover, the process of testing and experimentation by itself does not seem to be beneficial in the short term when it is not guided by an overarching theory. In fact, experimentation is usually not without costs (the parameter c in our model), since it involves investing managerial time and attention, and often also monetary commitments for large-scale testing, creation of MVPs, and so on. When guided by a

well-formulated theory, instead, startups might require less testing and therefore potentially incur fewer monetary costs and revenue delays.

CONCLUSION

This study provides novel and robust evidence that theory-based decision making under uncertainty significantly enhances firm performance. Across eight field experiments with early-stage startups spanning diverse economies and sectors, our findings indicate that entrepreneurs trained to formulate and test causal theories achieve superior outcomes in revenue growth and productivity, relative to those relying solely on iterative experimentation or operating without structured intervention. In essence, the capacity to theorize about cause and effect – not merely to collect data – emerges as a critical determinant of economic success. Our theoretical framework, supported by rigorous experimental evidence, thus reinforces the view that causal, theory-based reasoning should be considered a cornerstone of effective decision making in uncertain environments (Gilboa, Minardi, and Samuelson, 2020).

Looking ahead, several avenues merit exploration. Extending the analysis to broader contexts – including more mature firms – could help ascertain the generalizability of our results. While our study design involving eight RCTs across different countries and continents is a first step towards increasing external validity, scholars might be interested in the performance effects of causal theories in a variety of different contexts.

Moreover, integrating longer-term performance metrics and more granular evaluations of decision-making processes could provide deeper insights into the mechanisms through which theory-based reasoning translates into sustained competitive advantage. Future research might therefore combine our approach with complementary frameworks, such as behavioral or network models, to better understand how causal, theory-based decision-making interacts with other determinants of economic success.

In sum, our work underscores the transformative potential of coupling causal theories with experimentation in decision making under uncertainty, opening fertile ground for further theoreti-

cal refinement and empirical investigation.

ARNALDO CAMUFFO: Bocconi University and ION Management Science Lab, Milan, Italy. Email: arnaldo.camuffo@unibocconi.it

ALFONSO GAMBARDELLA: Bocconi University and ION Management Science Lab, Milan, Italy. Email: alfonso.gambardella@unibocconi.it

DIEGO JANNACE: Bocconi University and ION Management Science Lab, Milan, Italy. Email: diego.jannace@unibocconi.it

REFERENCES

Agarwal, Rajshree, Francesca Bacco, Arnaldo Camuffo, Andrea Coali, Alfonso Gambardella, Haji Msangi, Steven Sonka, Anna Temu, Betty Waized, and Audra Wormald (2025), "Does a theory-of-value add value? Evidence from a randomized control trial with Tanzanian entrepreneurs", *Organization Science* 36.(2) 2025, 601–625.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004), "How Much Should We Trust Differences-in-Differences Estimates?", *The Quarterly Journal of Economics* 119.(1) 2004, 249–275.

Bloom, Nicholas, Benn Eifert, Aprajit Mahajan, David McKenzie, and John Roberts (2013), "Does Management Matter? Evidence from India", *The Quarterly Journal of Economics* 128.(1) 2013, 1–51.

Bloom, Nicholas and John Van Reenen (2010), "Why Do Management Practices Differ across Firms and Countries?", *Journal of Economic Perspectives* 24.(1) 2010, 203–224.

Cagetti, Marco and Mariacristina De Nardi (2006), "Entrepreneurship, frictions, and wealth", *Journal of political Economy* 114.(5) 2006, 835–870.

Camuffo, Arnaldo, Alessandro Cordova, Alfonso Gambardella, and Chiara Spina (2020), "A Scientific Approach to Entrepreneurial Decision Making: Evidence from a Randomized Control Trial", *Management Science* 66.(2) 2020, 564–586.

Camuffo, Arnaldo, Alfonso Gambardella, Danilo Messinese, Elena Novelli, Emilio Paolucci, and Chiara Spina (2024), "A Scientific Approach to Entrepreneurial Decision-Making: Large-Scale Replication and Extension", *Strategic Management Journal* 45.(6) 2024, 1209–1237.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val (2025), "Fisher–Schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India", *Econometrica* 93.(4) 2025, 1121–1164.

Clarke, Damian, Joseph P Romano, and Michael Wolf (2020), "The Romano–Wolf Multiple-Hypothesis Correction in Stata", *The Stata Journal* 20.(4) 2020, 812–843.

Fisher, Ronald A (1935), *The Design of Experiments*, (Oliver & Boyd, Edinburgh).

Floyd, Eric and John A List (2016), "Using Field Experiments in Accounting and Finance", *Journal of Accounting Research* 54.(2) 2016, 437–475.

Frankel, Alexander and Emir Kamenica (2019), "Quantifying information and uncertainty", *American Economic Review* 109.(10) 2019, 3650–3680.

Gans, Joshua S (2023), "Experimental Choice and Disruptive Technologies", *Management Science* 69.(11) 2023, 7044–7058.

Gelman, Andrew and Guido W Imbens (2013), *Why Ask Why? Forward Causal Inference and Reverse Causal Questions*, NBER Working Paper 19614, National Bureau of Economic Research, 2013.

Gerber, Alan S and Donald P Green (2012), *Field Experiments: Design, Analysis, and Interpretation*, (W. W. Norton & Company, New York).

Ghanem, Dalia, Sarojini Hirshleifer, and Karen Ortiz-Becerra (2023), "Testing Attrition Bias in Field Experiments", *Journal of Human Resources* 58.(6) 2023, 1857–1892.

Gilboa, Itzhak, Stefania Minardi, and Larry Samuelson (2020), "Theories and cases in decisions under uncertainty", *Games and Economic Behavior* 123 2020, 22–40.

Gilboa, Itzhak and Larry Samuelson (2012), "Subjectivity in inductive inference", *Theoretical Economics* 7.(2) 2012, 183–215.

Global Entrepreneurship Monitor (2025), *GEM 2024/2025 global report: entrepreneurship Reality check*, tech. rep., Global Entrepreneurship Research Association (GERA): London, UK, 2025.

Heß, Simon (2017), "Randomization inference with Stata: A guide and software", *The Stata Journal* 17.(3) 2017, 630–651.

Hirano, Keisuke, Guido W Imbens, and Geert Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica* 71.(4) 2003, 1161–1189.

Holtz-Eakin, Douglas, David Joulfaian, and Harvey S Rosen (1994), "Sticking it out: Entrepreneurial survival and liquidity constraints", *Journal of Political economy* 102.(1) 1994, 53–75.

Imbens, Guido W (2020), "Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics", *Journal of Economic Literature* 58.(4) 2020, 1129–1179.

Imbens, Guido W and Donald B Rubin (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, (Cambridge University Press, New York).

John, Fitzgerald, Peter Gottschalk, and Robert Moffitt (1998), "An analysis of sample attrition in panel data: The michigan panel study of income dynamics", *Journal of Human Resources* 33.(2) 1998, 251–99.

Karni, Edi (2022), "A theory-based decision model", *Journal of Economic Theory* 201 2022, 105444.

Lee, David S. (2009), "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects", *The Review of Economic Studies* 76.(3) 2009, 1071–1102, (visited on 02/19/2025).

List, John A, Azeem M Shaikh, and Yang Xu (2019), "Multiple Hypothesis Testing in Experimental Economics", *Experimental Economics* 22 2019, 773–793.

Little, Roderick J and Donald B Rubin (2019), *Statistical Analysis with Missing Data (3rd ed.)* (John Wiley & Sons, Hoboken NJ).

Little, Roderick J and Linda Yau (1996), "Intent-to-treat analysis for longitudinal studies with drop-outs", *Biometrics* 1996, 1324–1333.

Moscarini, Giuseppe and Lones Smith (2001), "The Optimal Level of Experimentation", *Econometrica* 69.(6) 2001, 1629–1644.

Osterwalder, Alexander and Yves Pigneur (2010), *Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers*, vol. 1, (John Wiley & Sons, Hoboken NJ).

Thiel, Peter and Blake Masters (2014), *Zero to one: Notes on startups, or how to build the future*, (Crown Currency).

Young, Alwyn (2019), "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results", *The quarterly journal of economics* 134.(2) 2019, 557–598.

Zhong, Weijie (2022), "Optimal Dynamic Information Acquisition", *Econometrica* 90.(4) 2022, 1537–1582.

FIGURES

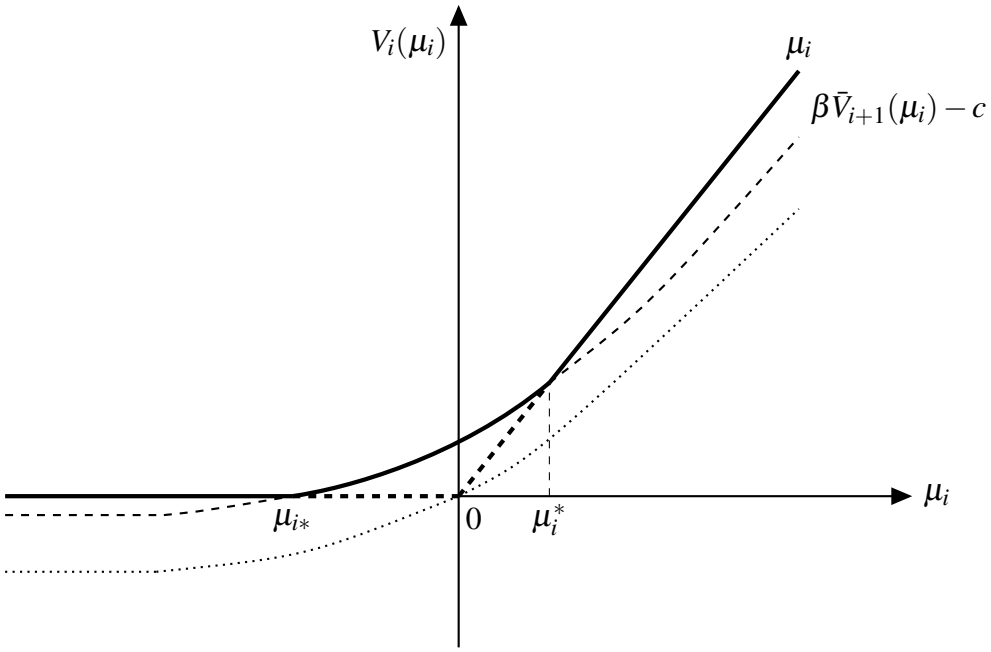


Figure I

This figure represents $V_i(p)$ – thick portion of the curves. The thresholds before and after which entrepreneurs terminate or commit are (μ_{i*}, μ_i^*). The threshold 0 denotes the discontinuous change of $v(\mu_i)$ in equation (1). The non-linear curves represent the net expected value of the experiment, and the thick portion of the upper curves denotes the experimentation zone. The dotted non-linear curve is the continuation value when the parameters β or c are such that there is no experimentation zone.

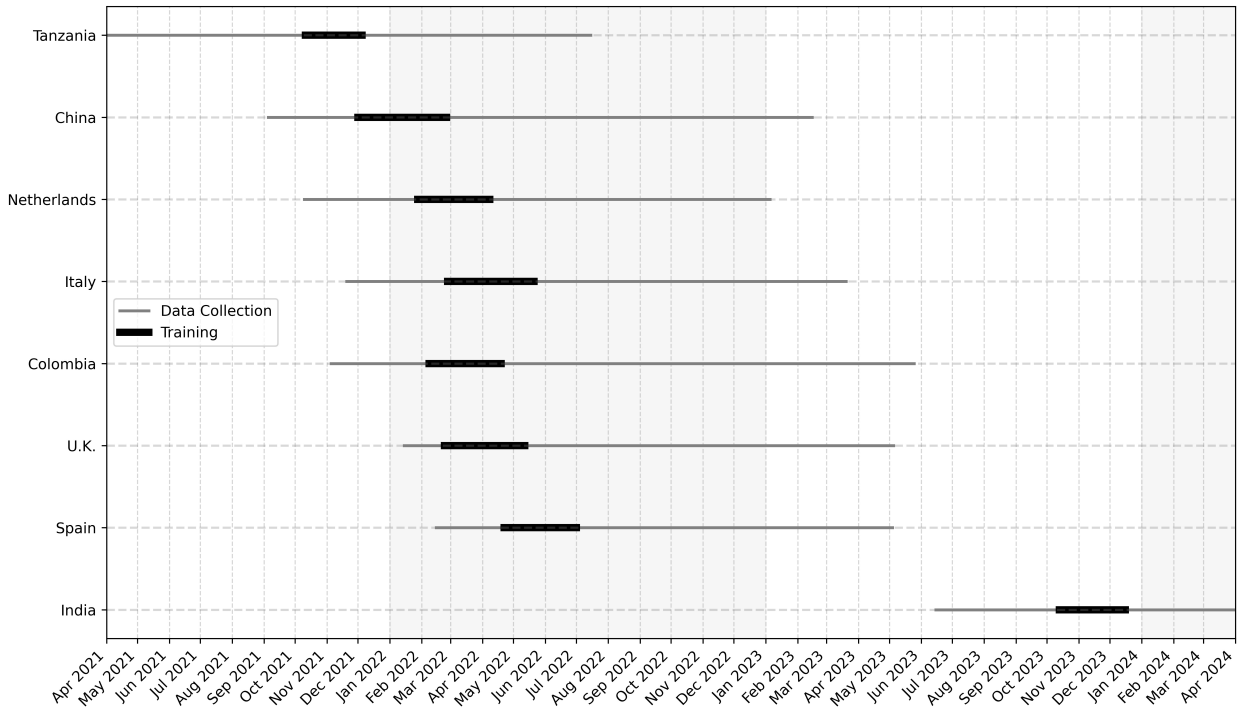


Figure II

This figure shows timelines for each of the eight RCTs that compose our study sample. Thin gray lines denote the data collection period, while thick black lines denote the training/treatment period.

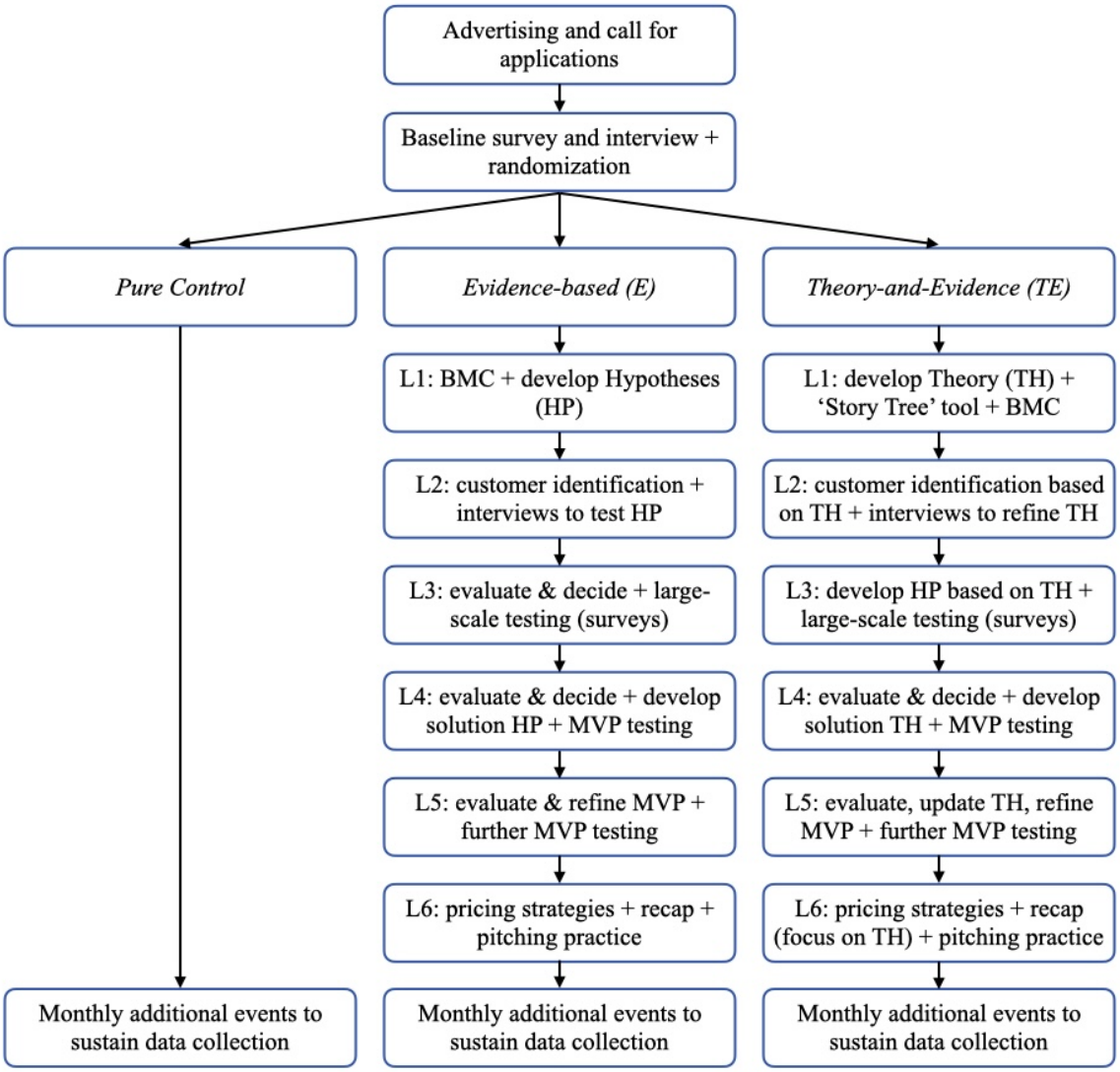


Figure III

This figure shows a schematic representation of the experimental design that was replicated in each RCT location. It also includes a broad overview of the key contents of each of the six four-hour treatment sessions. Appendix A then contains extensive information on the treatments.

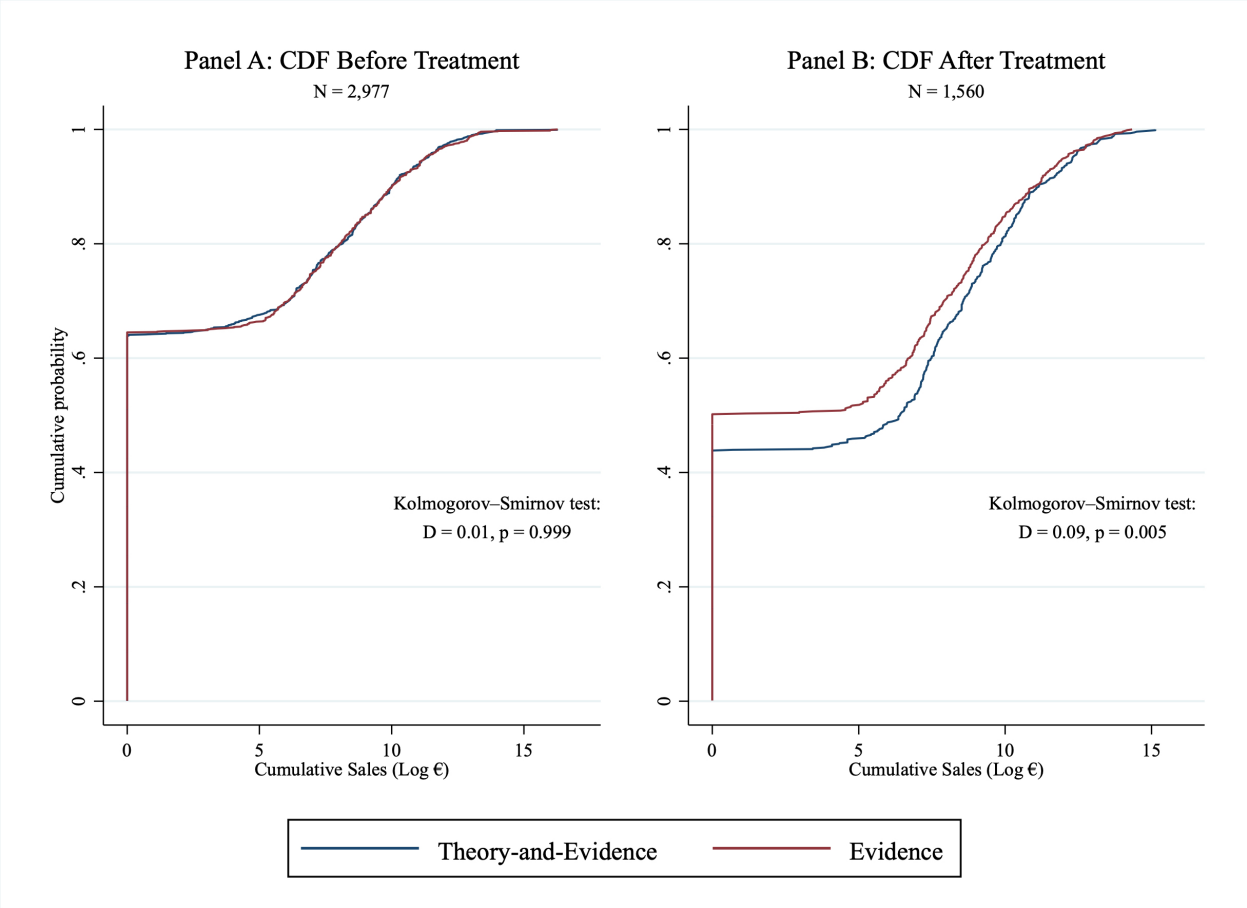


Figure IV

This figure depicts the cumulative distribution functions (CDFs) of the *Cumulative Sales (Log €)* outcome variable for *Theory-and-Evidence* and *Evidence* startups. Panel A plots the CDFs for the baseline and during-treatment periods, while Panel B plots the CDFs for the post-treatment periods, consistent with our difference-in-differences regressions. Both panels also report the results of two-sample Kolmogorov–Smirnov tests for equality of the distribution functions.

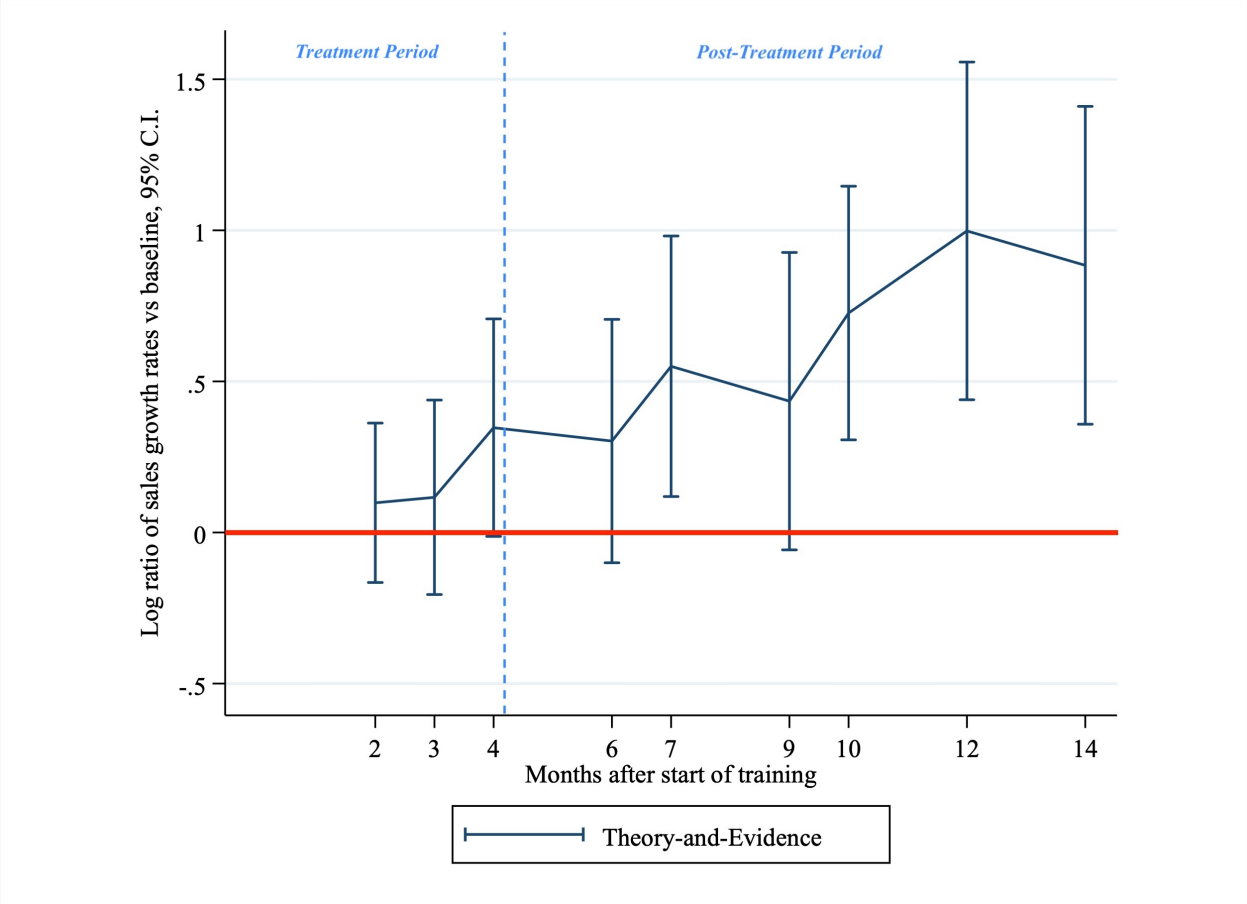


Figure V

This figure graphs the coefficients of a difference-in-differences regression of *Cumulative Sales (Log €)* on the subset of treated startups, using time and startup fixed effects, as well as standard errors clustered at the startup level. By interacting the treatment with the time variable (i.e., months since the start of the interventions), we are able to compare the growth rates of the two treatments over time. The blue curve in this figure represents the *Theory-and-Evidence* treatment, where the counterfactual (i.e., the thick red line at $y = 0$) is the *Evidence* group. Vertical bars represent 95% confidence intervals for the coefficients estimates.

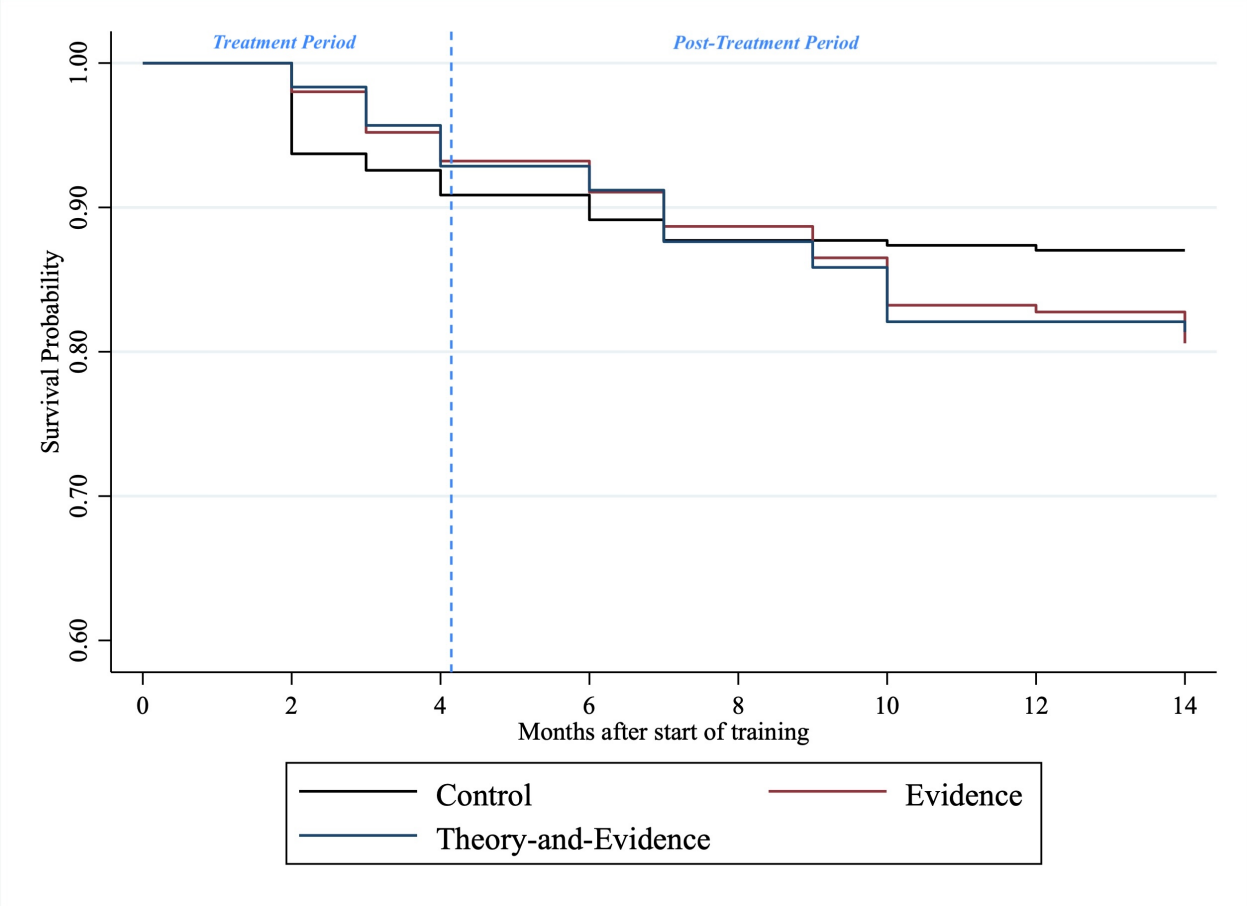


Figure VI

This figure shows Kaplan-Meier survival curves to present termination rates, plotted by treatment category. The failure event is represented by the binary *Termination* variable, which indicates the moment in time when a startup ceases its operations. As explained in the data section, this variable is based on a conservative assumption under which we only count the cases of confirmed termination. The number of post-baseline observations for this figure is $N = 6,867$. Log-rank tests of the difference in time to termination between conditions show weak support when comparing the *Control* group with the *E* one ($\chi^2 = 2.79, p = 0.095$) or with the *TE* one ($\chi^2 = 2.60, p = 0.107$). A comparison between the two treatment conditions indicates that they are not statistically different from each other ($\chi^2 = 0.00, p = 0.964$).

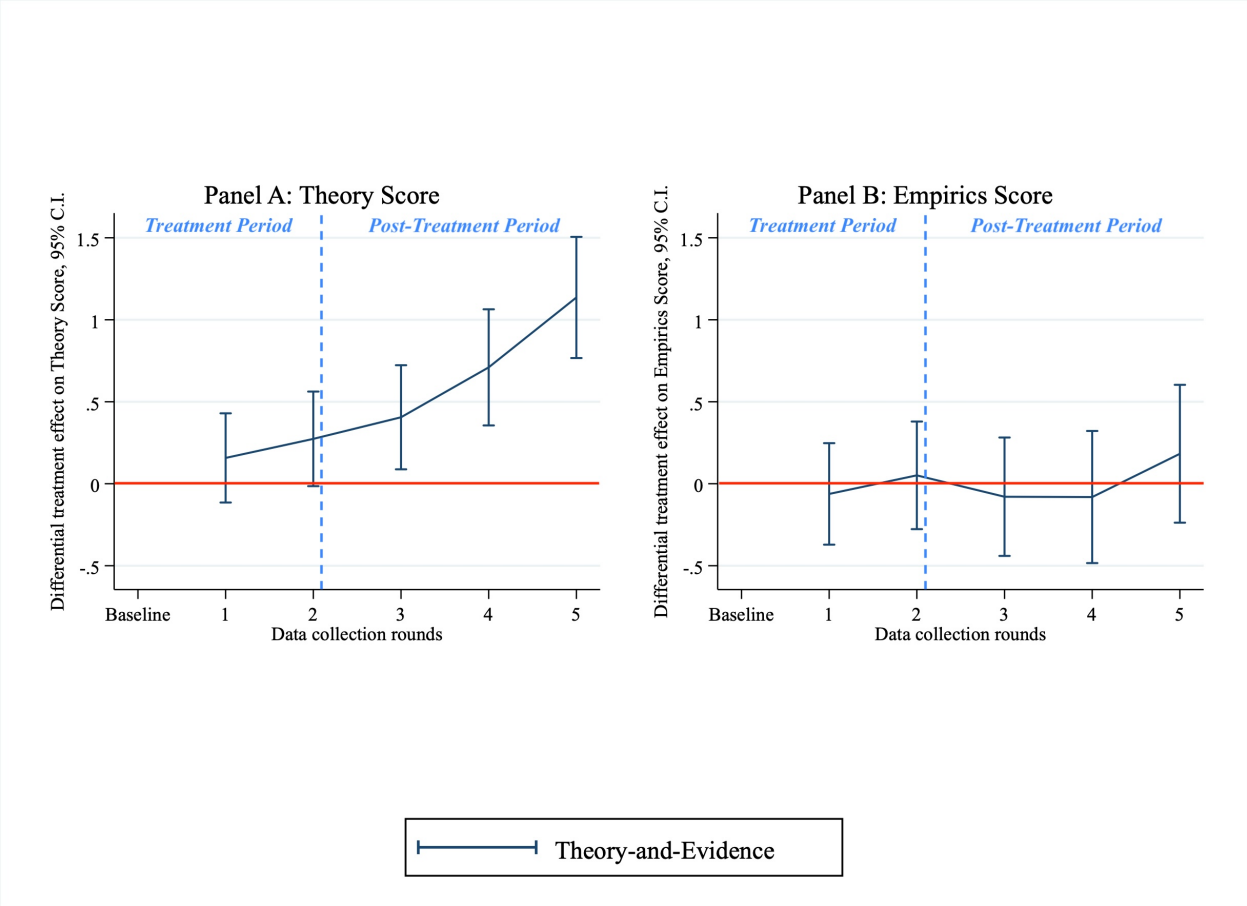


Figure VII

This figure plots the coefficients of difference-in-differences regressions of theory and experimentation scores on the subset of treated entrepreneurs, where the blue curves represent the *TE* treatment and the counterfactual (i.e., the thick red line at $y = 0$) is the *E* treatment. The outcome variable in Panel A is the startups' *Theory Score*, whereas in Panel B it is *Empirics Score*. Both regressions employ interview round and startup fixed effects, as well as standard errors clustered at the startup level. Vertical bars represent 95% confidence intervals for the coefficients estimates.

TABLES

TABLE I
Number of Startups by RCT

RCT #	RCT Country	RCT City (for training)	# Unique Startups	Percentage (%)
1	Tanzania	Dar-es-Salaam & Morogoro	202	12.98
2	China	Shanghai	167	10.73
3	Netherlands	Rotterdam	175	11.25
4	Italy	Milan	254	16.32
5	Colombia	Bogotá	207	13.30
6	U.K.	London	208	13.37
7	Spain	Barcelona	148	9.51
8	India	Chennai	195	12.53
Total			1,556	100.00

Notes. Number of unique startups by location of each RCT that composes our study sample. RCTs are listed in chronological order.

TABLE II
Variables and Definitions

Variable	Definition
Theory-and-Evidence	Indicates participants assigned to the Theory-and-Evidence-based treatment arm
Evidence	Indicates participants assigned to the Evidence-based treatment arm
Control (baseline category)	Indicates participants assigned to the Pure Control arm
Termination [s,i]	Indicates whether the startup ceases its activities and is no longer operating
Cumulative Sales (Log €) [s,i]	Startup cumulative sales reported in logged Euros *
Employees [s]	Number of founders, full-time, and part-time employees
Months of Operations [s]	Number of months that the startup has been operational for
Startup Sells a Product [s]	Dummy=1 if the startup only sells products
Startup Sells a Service [s]	Dummy=1 if the startup only sells services
Startup Sells Product + Service [s]	Dummy=1 if the startup sells both products and services
Startup is For-Profit [s]	Dummy=1 if the startup is for-profit (=0 if non-profit)
Perceived Prob. of Termination [s]	Perceived probability of terminating the business venture
Age of founder [s]	Age of the startup's main founder at baseline
Gender (=1 Female) [s]	Gender of the startup's main founder (0 if male, 1 if female)
Level of Education (1-3) [s]	Highest level of education of the startup's main founder at baseline. 1 for primary, 2 for secondary, 3 for tertiary
Econ or Business Degree [s]	Dummy=1 if the startup's main founder possesses a degree in economics or business-related subjects
Working Full Time [s]	Dummy=1 if the startup's main founder declares working full time for the startup
Hours Worked per Week [s]	Number of weekly hours dedicated to working on the startup by the main founder
Work Experience (Years) [s]	Total work experience in years of the startup's main founder
Development Stage (1-5) [i]	Startup's stage of development: from 1 (problem analysis) to 5 (stable sales)
Theory Score [i]	Extent to which the startup's business theory is clear, causal, based on evidence, and articulated in hierarchical components
Empirics Score [i]	Extent to which the startup collected appropriate evidence (e.g., tests, interviews, questionnaires, MVPs, ...), evaluated the collected data, and made decisions based on the resulting insights

Notes. * Sales values for RCT countries that do not adopt the Euro as a currency were converted to Euro using the appropriate average exchange rate for each local RCT currency and data collection period, taken from the European Central Bank's official website. [s] denotes variables collected through surveys; [i] denotes variables collected through interviews. The sales and termination measures, being the key performance metrics, were collected both through survey and interview at each data point to increase accuracy and reliability.

TABLE III
Summary Statistics - Full Sample

	N	Mean	Median	Std. Dev.	Min	Max
Cumulative Sales (Log €)	1,556	2.55	0.00	4.08	0	16.54
Startup Employees	1,556	4.49	2.00	9.35	1	100
Months of Operations	1,556	15.58	7.00	21.55	0	268
Startup Sells a Product	1,556	0.26	0.00	0.44	0	1
Startup Sells a Service	1,556	0.36	0.00	0.48	0	1
Startup Sells Product and Service	1,556	0.39	0.00	0.49	0	1
Startup is For-Profit	1,556	0.92	1.00	0.26	0	1
Perceived Prob. of Termination	1,556	17.26	7.00	24.87	0	100
Age of founder	1,556	33.15	32.00	9.33	18	72
Gender (=1 Female)	1,556	0.36	0.00	0.48	0	1
Level of Education (1-3)	1,556	2.84	3.00	0.38	1	3
Econ or Business Degree	1,556	0.31	0.00	0.46	0	1
Working Full Time	1,556	0.40	0.00	0.49	0	1
Hours Worked per Week	1,556	24.52	19.00	22.16	0	100
Work Experience (Years)	1,556	9.40	7.00	8.38	0	40
Development Stage (1-5)	1,554	2.52	2.00	1.66	1	5
Theory Score	1,554	3.44	3.60	0.95	1	5
Empirics Score	1,554	2.34	2.33	1.29	1	5

Notes. This table reports baseline summary statistics for the full sample of startups in our study, comprising all three treatment arms (i.e., Theory-and-Evidence, Evidence, Control) as well as all eight RCT locations. The high max value for *Months of Operations* is driven by a handful of outliers from the Tanzania and China RCTs, where also more established small enterprises were allowed to apply to our program with a new and innovative project. The last three variables display two missing observations because of two entrepreneurs who performed the baseline data collection questionnaire but not the interview with our Research Assistants due to personal reasons.

TABLE IV

Tests of balance across treatment categories

	(1) Control	(2) TE	(3) E	(1)-(2)	(1)-(3)	(2)-(3)
	Mean / (Std.Err.)			Pairwise t-test P-values		
Cumulative Sales (Log €)	2.94 (0.22)	2.36 (0.16)	2.50 (0.17)	0.12	0.44	0.59
Startup Employees	3.82 (0.37)	4.81 (0.43)	4.54 (0.38)	0.12	0.31	0.41
Months of Operations	15.35 (1.06)	14.28 (0.81)	17.01 (0.97)	0.46	0.31	0.03
Startup Sells a Product	0.23 (0.02)	0.26 (0.02)	0.26 (0.02)	0.55	0.32	0.76
Startup Sells a Service	0.38 (0.03)	0.37 (0.02)	0.33 (0.02)	0.91	0.24	0.10
Startup is For-Profit	0.93 (0.01)	0.92 (0.01)	0.92 (0.01)	0.62	0.77	0.80
Perceived Prob. of Termination	16.60 (1.25)	18.31 (1.06)	16.59 (0.99)	0.56	0.65	0.22
Age of founder	34.08 (0.47)	33.01 (0.38)	32.75 (0.39)	0.44	0.19	0.72
Gender (=1 Female)	0.39 (0.03)	0.38 (0.02)	0.33 (0.02)	0.68	0.35	0.09
Level of Education (1-3)	2.85 (0.02)	2.84 (0.02)	2.84 (0.02)	0.21	0.38	0.82
Econ or Business Degree	0.36 (0.03)	0.29 (0.02)	0.31 (0.02)	0.33	0.72	0.46
Working Full Time	0.35 (0.03)	0.41 (0.02)	0.43 (0.02)	0.66	0.36	0.49
Hours Worked per Week	22.50 (1.08)	24.33 (0.91)	25.87 (0.94)	0.46	0.06	0.19
Work Experience (Years)	10.20 (0.43)	9.31 (0.35)	9.03 (0.34)	0.53	0.20	0.66
Development Stage (1-5)	2.55 (0.09)	2.49 (0.07)	2.52 (0.07)	0.46	0.82	0.77
Theory Score	3.37 (0.05)	3.47 (0.04)	3.46 (0.04)	0.24	0.22	0.79
Empirics Score	2.23 (0.07)	2.42 (0.05)	2.33 (0.05)	0.23	0.88	0.25
F-test of joint significance (F-stat)				0.99	0.99	1.12
Observations	350	602	604	952	954	1,206

Notes. This table reports the distribution of the baseline covariates of our study, split by treatment category, as well as tests of balance. Columns (1), (2), and (3) report the mean values and standard errors (in parentheses) for *Control*, *Theory-and-Evidence-based*, and *Evidence-based* startups, respectively. The last three columns show the p-values of the pairwise t-tests for each variable and treatment pair. At the bottom of the table, we also include a row with the F-tests for joint significance across all balance variables for each test pair, as well as the number of observations for each of the six columns. This table was produced with Stata 18, using the 'ietoolkit' package from the World Bank. All tests include standard errors clustered at the startup level, and fixed effects for each of our eight RCT locations because the randomization was performed within each location separately.

TABLE V
Revenue Analyses

	Total Sales (Log €)		Sales Dummy	Total Sales (Log €)
	(1)	(2)	(3)	(4)
Theory-and-Evidence × Post	0.53	0.58	0.04	0.25
	(0.16)	(0.16)	(0.02)	(0.11)
Evidence × Post	0.05	0.22	-0.00	0.05
	(0.14)	(0.15)	(0.02)	(0.09)
Constant	3.69	2.98	0.43	8.89
	(0.04)	(0.05)	(0.01)	(0.03)
$\beta_{TE} = \beta_E$ (p-value)	0.002	0.005	0.015	0.042
Observations	5,270	3,112	5,270	1,827
Startup FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Estimation Panel	Full	2-period	Full	Nonzero sales

Notes. This table reports difference-in-differences analyses of the effect of the treatments on startup revenue. For Columns (1), (2), and (4), the outcome is a measure of cumulative sales reported in logged Euros (€). For Column (3), the outcome is a binary dependent variable indicating whether the startup generates positive sales. Column (1) reports OLS difference-in-differences estimations using the entire panel data set. Column (2) estimates the same regression but using only the first (i.e., pre-treatment) and last available data points in our panel to eliminate concerns of serial correlation across periods biasing our standard errors. Column (2) estimates the probability of generating positive sales (full sample). Column (4) estimates the same model as Column (1) but only on the subsample of entrepreneurs who already had positive sales at baseline, to reduce skewness in the data. Each column also reports the two-sided p-value of Wald tests for the equality of the two treatment coefficients. All models include startup and time fixed effects, which cause singleton observations to be dropped. Standard errors (in parentheses) are clustered at the startup level in all columns.

TABLE VI
Employment and Productivity Analyses

	Employees	Log(Sales / Employees)	TFP
	(1)	(2)	(3)
Theory-and-Evidence × Post	0.20	0.62	0.53
	(0.24)	(0.18)	(0.16)
Evidence × Post	0.21	0.11	0.13
	(0.26)	(0.16)	(0.15)
Constant	4.91	3.25	0.01
	(0.06)	(0.04)	(0.10)
$\beta_{TE} = \beta_E$ (p-value)	0.972	0.004	0.008
Observations	5,104	5,104	5,104
Startup FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Estimation Panel	Full	Full	Full

Notes. This table reports difference-in-differences analyses of the effect of the treatments on employment and productivity. For Column (1), the employment outcome is the count of the total number of people working for the startup at each data collection point. For Column (2), the productivity outcome is the log of the ratio between startup total sales (in Euro) and the total number of people working for the startup. For Column (3), the productivity outcome is a proxy estimation of Total Factor Productivity (TFP), computed as the residuals from an OLS regression of Log Cumulative Sales on Log Employees (controlling for time and RCT location fixed effects). Each column also reports the two-sided p-value of Wald tests for the equality of the two treatment coefficients. All models include startup and time fixed effects, which cause singleton observations to be dropped. Standard errors (in parentheses) are clustered at the startup level in columns (1) and (2), while Column (3) reports bootstrap standard errors (200 reps) given the usage of residuals. The number of observations is lower than in Table V since the employment variable was only collected through surveys, whereas the sales measure was collected both via surveys and interviews at each data point, therefore resulting in slightly fewer missing values.

TABLE VII
Termination Analyses

	Termination Dummy			Termination Month
	(1)	(2)	(3)	(4)
Theory-and-Evidence	0.05	0.05	1.31	-2.84
	(0.02)	(0.02)	(0.24)	(1.39)
Evidence	0.05	0.06	1.32	-2.82
	(0.02)	(0.02)	(0.24)	(1.38)
Constant	0.19			31.38
	(0.04)			(1.87)
Observations	1,556	1,556	6,867	1,556
RCT Controls	Yes	Yes	Yes	Yes
Estimation	OLS	Probit	Cox HR	Tobit

Notes. This table reports analyses of termination (i.e., startup closure) for all startups in our sample. In the first three columns, the outcome is a binary variable indicating whether each startup ceased activities (i.e., terminated) during the study's time frame. Column (1) reports results from an OLS regression of the (linear) probability of terminating by treatment condition. Column (2) reports an analogous model estimated using Probit average marginal effects for robustness. Column (3) instead reports the output of a Cox proportional hazards regression predicting the hazard ratios of startup termination. The number of observations for this model is higher because it is estimated leveraging the full time-series information from our panel data set. Column (4) reports the results of a Tobit model that estimates the time to termination (in months since the start of training), accounting for the right-censoring of our data. All models include controls for RCT location and standard errors (in parentheses) clustered at the startup level.

TABLE VIII

Performance Regressions with Theory and Experimentation Scores

	Total Sales (Log €)	Employees	Log(Sales/Employees)
	(1)	(2)	(3)
Theory Score	0.85	0.24	0.83
	(0.40)	(0.48)	(0.41)
Empirics Score	-0.76	-0.19	-0.71
	(0.44)	(0.54)	(0.44)
Observations	4,807	4,723	4,723
Startup FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
First-stage tests			
Theory Score F-statistic	52.88	50.87	50.87
Empirics Score F-statistic	50.50	48.80	48.80
K-P rk LM test (p-value)	0.008	0.006	0.006
Cragg-Donald Wald F statistic	15.13	15.37	15.37

Notes. This table reports second-stage regressions examining the relationship between theory/empirics scores and startup performance, exploiting experimental variation induced by treatment assignment. Results should be interpreted as descriptive and mechanism-consistent rather than as causal instrumental-variable estimates. For Column (1), the revenue outcome is the measure of total sales reported in logged Euros (€). For Column (2), the employment outcome is the count of the total number of people working for the startup at each data collection point. For Column (3), the productivity outcome is the log of the ratio between startup total sales (in Euros) and the total number of people working for the startup. F-statistics are reported in each column for the regressions linking treatment assignment to the intermediate theory and empirics scores. Each column also reports standard first-stage diagnostic statistics (Kleibergen-Paap rk LM p-value and Cragg-Donald Wald F statistic), which are provided for completeness but should not be interpreted as validating a causal exclusion restriction. Complete regressions linking treatment assignment to theory and empirics scores are reported in Table B.I of Appendix B.2. All models include startup and time fixed effects, which cause singleton observations to be dropped. Standard errors (in parentheses) are clustered at the startup level in all columns.

APPENDIX A: DETAILED DESCRIPTION OF THE TREATMENTS

As described in Section 3.2, our experimental design revolved around comparing the effects of two treatments: *Evidence-based (E)* vs. *Theory-and-Evidence-based (TE)*. In the *E* condition, startups were encouraged to focus on experimentation: rapid validation cycles based on collecting information to test one or more components of a business model through customer interviews, large-scale surveys, A/B tests, and minimum viable product (MVP) trials. By contrast, in the *TE* condition entrepreneurs were required to anchor their decision making – including experimentation as seen for the *E* group – within a structured causal theory. These theories acted as the fundamental building block to which every treatment session and content was linked. Both treatment groups attended the same number of training hours, issued by the same instructors, in the same locations, and on consecutive days (or at different times on the same day).

A.1 Evidence-based Treatment

The *E* treatment teaches participants how to iteratively test and rapidly validate business hypotheses through data experimentation. In principle, a rigorous treatment would imply instructing participants about optimal information acquisition (Frankel and Kamenica, 2019; Zhong, 2022; Moscarini and Smith, 2001; Gans, 2023) including topics like experiments' efficiency, experimental design and experimental strategies. However, in order to make the treatment easily understandable and actionable to participants with a wide range of backgrounds, we resorted to translating optimal information acquisition strategies into intuitive and easy-to-use conceptual tools (like structured interviewing, MVPs and A/B tests, which we describe below) so that all participants, across the eight RCTs, could easily understand and learn how to efficiently acquire *additional* information about their business ideas. Participants in this condition are encouraged to systematically validate customer needs, develop MVPs to test solutions, and refine their business models based on experimental feedback. The idea is that experimentation – eliciting novel information – facilitates continuous learning, allowing entrepreneurs to revise, pivot, or abandon ideas based on empirical evidence. Experimentation enables entrepreneurs to refine their value propositions

through an adaptive, data-driven process. More specifically, this group of entrepreneurs is taught to validate their business ideas through two sequential processes. 'Problem validation' involves the assessment of whether a meaningful problem exists and its magnitude. 'Offer/solution validation' tests the attractiveness of the proposed solution to the target market, as well as evaluating its feasibility and scalability. The *E* treatment grounds information acquisition through experiments on structured, widely used frameworks such as the "business model canvas" (BMC) (originally introduced by Osterwalder and Pigneur, 2010), customer profiling ('customer personas'), and experimental techniques like A/B testing and MVP prototyping.

The curriculum of the *E* treatment comprises the following six training sessions:

1. Introduction to common entrepreneurial mistakes, development of a BMC to describe the startup's strategy, quick and continuous validation by building hypotheses based on the BMC cells.

2. Identification and mapping of early adopters through 'customer personas', understanding of the 'customer process' and relevant 'pain points', planning of customer interviews to test hypotheses.

3. Analysis of interview responses, validation of hypotheses to update the BMC, iteration by refining or developing new hypotheses and testing them through large-scale surveys and questionnaires.

4. Analysis of survey responses, validation of hypotheses to update the BMC, decision on whether to keep pursuing the project vs. refining it (i.e., 'pivoting') vs. terminating it. If the decision is to continue to pursue or pivot, shift from problem to offer/solution, followed by the creation of solutions (e.g., prototypes) to test through the development of a MVP.

5. Analysis of MVP test results, validation of hypotheses to update the BMC and the solution, new iteration of the pursue/refine/terminate decision. This cycle is repeated running comparative A/B tests of the the MVP or prototype features.

6. Overview of key pricing strategies for the startups' products or services, overall recap of the course and of the test-evaluate-decide cycle, final 'pitching' session to practice presentations to investors.

All sessions were complemented by homework (e.g., exercises and applications) to reinforce the absorption and implementation of the relevant concepts.

A.2 Theory-and-Evidence-based Treatment

The *TE* treatment includes the same content as the *E* treatment (75% of training hours) with the difference that it also includes content on how to formulate causal theories to guide testing and, more generally, entrepreneurial decision making. A theory clarifies why a business idea should succeed by establishing causal relationships among key elements of a problem. Causal reasoning allows one to make explicit not only the *effect* (a target future state, like the reduction of restaurant food waste in large cities) but also the *causes* (the events and hypothetical actions that might make the target future state more likely to occur, like the development of an app for relocating and reassigning food). In order to make the treatment easily understandable and actionable to participants with a wide range of backgrounds, we resorted to translating theory-based decision making logic into an intuitive and easy-to-use conceptual tool (the "story tree" or "theory tree", which we describe below) so that all participants across the eight RCTs could easily understand and learn how to develop causal theories of their business ideas. Entrepreneurs in this experimental condition were therefore taught to: i) identify critical elements of the problem or solution (define the consideration set); ii) causally connect such elements specifying why and with respect to what (counterfactual thinking); iii) develop beliefs (i.e., subjective probabilities) about such causal theory, and iv) test their theories through experimentation and update their beliefs. In practical terms, causal theorizing is embodied in the *TE* treatment through the 'story tree' canvas (introduced by Agarwal et al., 2025, see Figure A.1). This tool nudges entrepreneurs to do three things. First, it encourages them to clarify what problem or business opportunity they are focusing on or envisioning, such as a new market, a more efficient solution to an existing customer problem, or a new way to monetize an

existing business. This helps them to clearly articulate a problem or target future state. Second, entrepreneurs are instructed to identify the most important elements that are believed to be crucial antecedents for achieving their desired future goal. Entrepreneurs were trained to rank-order and prioritize these elements based on their perceived importance. Third, entrepreneurs are prompted to ask themselves *why* and identify what will make the target future state more (or less) likely to occur. The key goal is to cue participants' causal reasoning and help them identify causes that affect the likelihood of occurrence of the target future state (both directly and indirectly – e.g., longer causal chains). Within these causal models, participants are also nudged to ask *what can I do about it*, that is to identify their hypothetical interventions (e.g., a new technology, product, service or business model). Throughout the *TE* training treatment, entrepreneurs also learn to articulate their beliefs about both their probability of success (how likely is the target future under their causal theory) and the causal theory itself (how confident they are about their 'story tree' and causal connections. Overall, the 'story tree' was designed as a simple, intuitive, and effective tool to nudge participants to apply theory-based decision making using a visual representation similar to Directed Acyclic Graphs (DAGs) (Imbens, 2020).

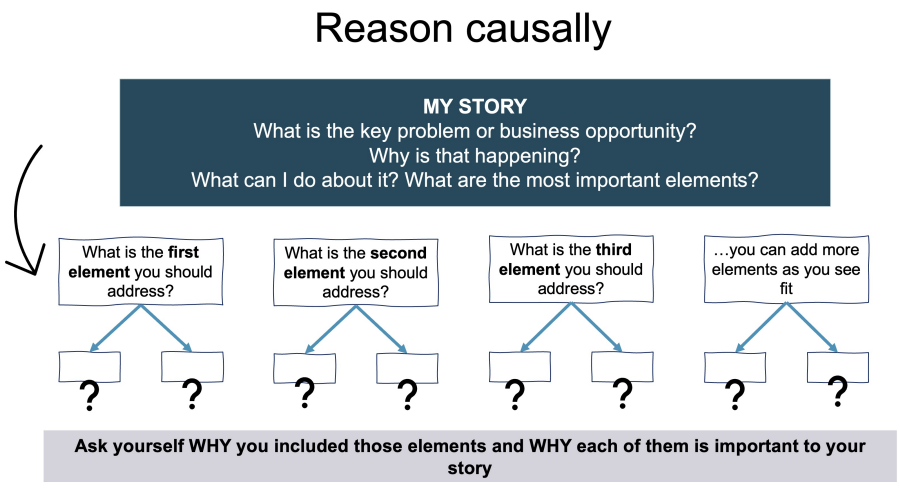


Figure A.1
The 'Story Tree' canvas nudging causal reasoning in theory-building, reproduced from Session 1 training slides of the *TE* treatment.

The curriculum of the *TE* treatment comprises the following six training sessions:

1. Introduction to common entrepreneurial mistakes, introduction to the scientific approach and to the 'story tree' canvas to guide the startup's strategy. Introduction to the BMC, which has to be linked to the theory tree.

2. Development of a causal theory for the business, understanding and framing of the key problem to solve, mapping of the 'customer process', understanding of assumptions in the theory about target customers and identification of 'customer personas'. Planning of exploratory interviews with target customers to validate the theory's core assumptions and causal links.

3. Usage of interview insights to update the story tree, breaking down the causal theory into testable hypotheses, and testing them through large-scale surveys and questionnaires. Formulation of prior beliefs, counterfactuals, and thresholds for falsifiability before collecting data.

4. Analysis of survey responses, validation of hypotheses to update the theory, decision on whether to keep pursuing the project vs. refining it (i.e., 'pivoting') vs. terminating it. If the decision is pursue or pivot, shift from problem to offer/solution, followed by the creation of solution hypotheses based on the theory. Development of a MVP based on a sub-theory of which components to include and why.

5. Analysis of MVP test results, validation of hypotheses to update the theory and the solution, new iteration of the scientific approach cycle by running comparative A/B feature tests through the MVP or prototype.

6. Overview of key pricing strategies for the startups' products or services, overall recap of the course and of the scientific approach (i.e., theory-based) cycle, final 'pitching' session to practice presentations to investors.

All sessions were followed by at-home exercises to reinforce the understanding of and apply the relevant concepts.

APPENDIX B: MEASURING 'THEORY' AND 'EMPIRICS'

This section outlines the methodology used to assess treatment absorption (compliance), i.e. the extent to which study participants applied causal theories and experimentation in their entrepreneurial decision making. This data collection effort was specifically designed to allow the exploration and testing of the hypothesized theoretical mechanisms underlying the *E* and *TE* treatments.

B.1 Data collection

Throughout the study, our research assistants (RAs) conducted structured phone interviews with randomly assigned participants at predetermined intervals. These interviews followed a standardized script featuring open-ended questions designed to capture the application of both causal theories and experimentation (as components of the 'scientific' approach proposed in Camuffo et al., 2020 and Camuffo et al., 2024). This data collection approach is similar to that employed by Bloom and Van Reenen, 2010 in the 'World Management Survey', where qualitative interview responses are systematically coded to identify behavioral patterns. In our study, RAs regularly interviewed the participants and analyzed their responses to determine the extent to which they developed, tested and updated causal theories and beliefs. To minimize bias in responses, participants were unaware of the criteria used for evaluation.

The use of causal theories and experimentation was measured through a behavioral coding system, with scores ranging from 1 (lowest score) to 5 (highest score) across five key dimensions: (i) Theory, (ii) Hypotheses, (iii) Evidence (i.e., tests), (iv) Evaluation, and (v) Decision. Each dimension was treated as a multidimensional construct, encompassing various characteristics and sub-components. For instance, causal theory application comprises clarity, causality, plausibility and articulation. RAs coded the interviews for the frequency and intensity of occurrence of these sub-components. To ensure reliability in assessment, all RAs underwent intensive training before participating in data collection and interview coding. This training involved multiple training and calibration sessions about how to conduct and code interviews according to the protocol. It also included practice interviews and scoring exercises to establish a uniform understanding of the data

collection and coding protocol. By standardizing interviewing, coding, and evaluation criteria, we minimized discrepancies in scoring across research assistants and ensured consistency in data interpretation.

Our *Theory Score* variable is an index comprising the evaluation of the following components that aim to measure the extent of theory-based decision making:

- Clarity: The extent to which the theory and its underlying logic is understandable.
- Causality: The extent to which the theory is based on causal logic and on causally connecting its conceptual components.
- Plausibility: The extent to which the business idea is grounded on a (observable) problem or target future state.
- Articulation: The extent to which the theory clearly articulates the causes.
- Hierarchy: The extent to which causes and components are weighted, with some being identified as more important than others.

The *Empirics Score* variable is an index comprising the evaluation of the following components that aim to capture the collection of evidence through experiments:

- 'Evidence' block: five components which measure information acquisition (test coherence, validity, representativeness, rigor, and causality).
- 'Evaluation' block: five components which measure data analysis and assessment (alignment with theory, rigor, prior expectations, and updating of priors).
- 'Decision' block: two components which measure the extent to which decisions logically derive from the results of the tests, and the consideration of alternative options.

RAs scored each of the components described above on a scale from 1 (lowest score) to 5 (highest score) for each interview at each data point[12]. For entrepreneurs who reported not having

[12]The full codebook and documents containing RA instructions and training sessions are available upon request.

run any tests, the corresponding scores were automatically set to the lowest value of 1 for that period. At each data collection period, scores were then averaged together – following the 'theory' vs. 'empirics' split – to smooth out potential outliers in single components. Finally, the two resulting indexes were then cumulated over time to generate the *Theory Score* and *Empirics Score* variables described in Table II. This accumulation over time was necessary for two key reasons: (i) to mirror the cumulative nature of our performance variables, and (ii) because the interview questions were designed to assess the causal theories and testing specifically carried out between two consecutive data collection interviews.

B.2 First-stage effects: Treatment validity

To reinforce the interpretation of our results described in Section 5, we ensured that our treatments were valid. Following our theoretical framework and experimental design, we are interested in verifying that the *E* and *TE* treatments increased evidence collection through experiments and causal theories, respectively. More specifically, we wanted to verify that the *E* and *TE* treatments caused a significant first-stage effect on the *Empirics Score* and on the *Theory Score*, respectively. We therefore estimate a difference-in-differences regression based on Equation 6, where the dependent variables Y_{ijt} are the *Theory* (Columns (1) and (2)) and *Empirics* scores (Columns (3) and (4)), and DD_{ijt} indicates whether the startup i has taken treatment j at time t.

TABLE B.I
First-stage effects: Treatment validity

	Theory Score		Empirics Score	
	(1)	(2)	(3)	(4)
Theory-and-Evidence × Post	2.08	1.84	1.70	1.83
	(0.21)	(0.20)	(0.20)	(0.20)
Evidence × Post	1.45	1.30	1.66	1.74
	(0.20)	(0.19)	(0.20)	(0.20)
Constant	9.40	9.45	6.30	6.27
	(0.05)	(0.05)	(0.05)	(0.05)
$\beta_{TE} = \beta_E$ (p-value)	0.002	0.003	0.848	0.693
Observations	4,830	4,830	4,830	4,830
Startup FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Coder FE	No	Yes	No	Yes

Notes. This table reports difference-in-differences analyses of the effect of the treatments on different aspects of "scientific" reasoning. These models represent the first stages of our 2SLS regression analyses. For Columns (1) and (2), the outcome is the participants' *Theory Score*, which measures the extent to which each entrepreneur's business theory is clear, causal, based on evidence, and articulated in different hierarchical components. For Columns (3) and (4), the outcome is the participants' *Empirics Score*, which measures the extent to which each entrepreneur collected appropriate evidence (e.g., through tests, interviews, questionnaires, MVPs, etc.), evaluated the collected data, and took decisions based on the resulting insights. Each column also reports the two-sided p-value of Wald tests for the equality of the two treatment coefficients. All models include startup and time fixed effects, which cause singleton observations to be dropped. Columns (2) and (4) further add coder fixed effects to account for potential individual differences in the ways that RAs conduct and codify interviews. Standard errors (in parentheses) are clustered at the startup level in all columns.

Estimates in Column (1) of Table B.I indicate that the *TE* treatment caused an average increase of the *Theory Score* by 2.08 points ($\beta_{TE} = 2.08, p = 0.000$) compared to the control group, with the *E* treatment's increase being 1.45 points ($\beta_E = 1.45, p = 0.000$). Wald tests indicate that the two coefficients are statistically different from each other ($p = 0.002$). When tabulating the post-treatment means, the mean *Theory Score* of the control startups is 14.21, versus 16.13 for *E* startups and 17.00 for *TE* ones. Two-sample t-tests indicate that the *TE* group's post-treatment mean is significantly higher than the control group ($p = 0.000$) and the *E* group ($p = 0.000$). These results confirm that our theory-based treatment was valid in causing a measurable increase in the

participant's theory formulation and quality, as compared to the other two experimental groups. The increase generated by the evidence-based treatment can also be explained by the usage of tools such as the Business Model Canvas and the formulation of hypotheses, which might indirectly be reflected in the theory index.

The estimates in Column (3) of Table B.I then show that both the *TE* treatment ($\beta_{TE} = 1.70, p = 0.000$) and the *E* one ($\beta_E = 1.66, p = 0.000$) caused a significant increase in the treated participants' *Empirics Score*. In line with our experimental design, these two coefficients are not statistically different from each other ($p = 0.848$). Post-treatment means are equal to 8.23 for the control group, 11.47 for the *E* one, and 11.44 for the *TE* group. Two-sample t-tests confirm that the means of both treatment groups are higher than that of the control group ($p = 0.000$), while they are not statistically different from each other ($p = 0.884$).

Although Columns (1) and (2) report the same first stage specifications as those run in Table VIII, we strive to provide robustness tests by also including RA fixed effects to account for potential systematic differences in coding. Whereas all RAs underwent a lengthy training and testing period in all RCT locations and followed a standardized codebook, there is still a margin for individual interpretation when assigning interview scores. Columns (2) and (4) therefore repeat the analyses of Columns (1) and (3), respectively, with the addition of coder fixed effects. Estimates presented in Column (2) and (4) are largely aligned with previous results, confirming the validity of our treatments and codebook.

B.3 Robustness tests via LLM evaluations

As described above, we collected data on the theories and tests of the startups through interviews conducted by randomly assigned RAs. Due to logistical and budgetary constraints, however, only one RA interviewed a startup at any given data collection round. The detailed interview script and codebook, along with a thorough process of selection and training of the RAs, helped to minimize the risk of spurious evaluations.

To get further reassurance about the accuracy and reliability of our theory and empirics

scores, we perform generative-AI-assisted coding and scoring of the interviews in order to get an alternative evaluation of the *Theory Score* and *Empirics Score*. As highlighted in the paper, the participants' interviews were systematically recorded and organized into RCT-specific directories. They were then converted into uniform .mp3 audio format, and transferred to Bocconi University's high-performance computing server, where an offline implementation of OpenAI's 'Whisper' neural net generated accurate text transcriptions, preserving the local interview language[13]. We then submitted this corpus of text data to OpenAI's 'GPT' API[14] and to Google's latest 'Gemini' API, employing a series of precise, multi-layered, and structured prompts, where the LLMs were instructed to act as a RA and were given the same code book and interview coding instructions provided to RAs. Overall, we were able to recover, transcribe, and re-evaluate 93.2% of the original 5,220 interviews across our eight RCTs. The remaining ones were missing due to lack of recordings, corrupted files, or minor RA errors (e.g., unreconcilable file names that did not allow for unquestionable identification of the interviewee and/or the interview round).

Starting with OpenAI's GPT, each interview transcript was evaluated twice by the LLM, to reduce potential randomness. Mean differences between the two GPT AI evaluations were -0.01 and 0.02 for *Theory Score* and *Empirics Score*, respectively, whereas the median differences were equal to 0 in both cases, indicating that further iterations were not necessary. We then averaged the two LLM evaluations for each interview and compared them with our human scores. Figure B.1 compares the distributions of the RA- and LLM-generated scores. Panel A reports the CDFs of the theory scores, highlighting that the GPT AI model is slightly more conservative in assigning high scores compared to RAs. The mean difference between the two scores is 0.51 points (that is, 0.27 standard deviations), while the median is reassuringly equal to 0. Panel B then plots the CDFs of the testing and empirics scores, showing an overall very close overlap between the two curves. The mean difference between RA and LLM *Empirics Scores* is equal to 0.07 (i.e., 0.05

[13] Across our eight RCT locations, interviews were performed in the following languages: Chinese, Dutch, English, Italian, Spanish, Swahili. We verified the AI models' ability to both accurately transcribe each of these languages, as well as to understand them when 'reasoning'.

[14] We employed the o3-mini model, which was the latest available at the time of writing. We used an enterprise account to preserve privacy and confidentiality of the data.

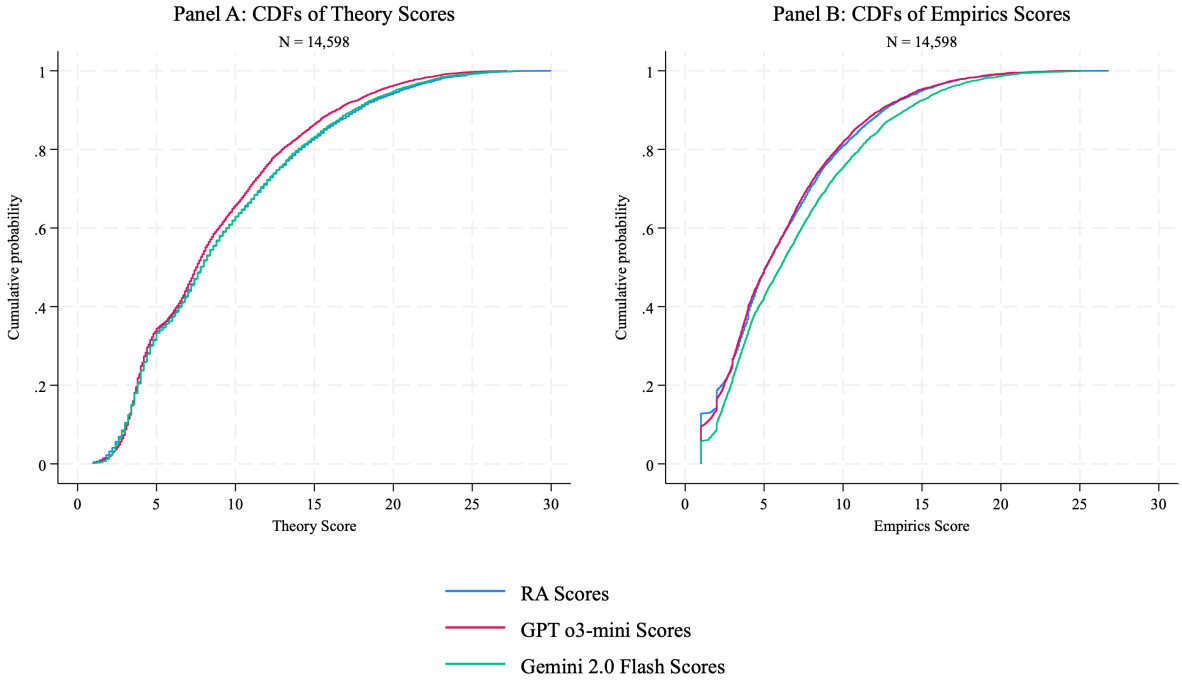


Figure B.1

This figure graphs the cumulative distribution functions (CDFs) of the *Theory Score* (Panel A) and *Evidence Score* (Panel B) variables, comparing the human RA scores to the LLM-generated ones (GPT o3-mini and Gemini 2.0 Flash, the latest suitable models available at the time of writing).

standard deviations) and the median difference is 0. For each panel, Kolmogorov-Smirnov tests comparing the maximum absolute difference between the curves reject the null hypothesis that the distributions are equal, although the differences are quite small (Panel A: $D = 0.05, p = 0.000$; Panel B: $D = 0.03, p = 0.013$).

To further ascertain the reliability of our RA-coded measures, we also follow the same AI-scoring procedure highlighted above, employing a second LLM model for robustness. For this second test, we employ Google's 'Gemini Flash 2.0 ' API, which was the latest version available at the time of writing. Figure B.1 (green curves) also reports the results of this second LLM (average of two repeats, as above). The comparison among CDFs in Panel A highlights a close alignment between this second AI model and the RAs when evaluating *Theory Score*. However, the model tends to differ on the evaluation of the *Empirics Score* (Panel B), with the human RAs being on average more conservative.

TABLE B.II
2SLS: Performance Analyses - LLM Robustness

	Total Sales (Log €)	Employees	Log(Sales/Employees)
	(1)	(2)	(3)
Theory Score (3 sources)	1.27	0.52	1.27
	(0.69)	(0.72)	(0.71)
Empirics Score (3 sources)	-1.23	-0.44	-1.21
	(0.77)	(0.83)	(0.78)
Observations	4,461	4,382	4,382
Startup FE	Yes	Yes	Yes
Time FE	Yes	Yes	Yes
Theory Score F-statistic	53.34	51.22	50.87
Empirics Score F-statistic	48.82	46.57	48.80
Kleibergen-Paap rk LM test (p-val)	0.029	0.025	0.025
Cragg-Donald Wald F statistic	10.10	10.27	10.27

Notes. This table repeats the same analyses as Table VIII, using the average between evaluations from human RAs and the two LLM models for compiling the *Theory Score* and *Empirics Score* variables. F-statistics are reported in each column for the two first stage regressions. Each column also reports the appropriate first-stage statistics for the underidentification test (Kleibergen-Paap rk LM p-value) and for the weak identification test (Cragg-Donald Wald F statistic). All tests in all columns indicate that our first-stage regressions are properly identified. All models include startup and time fixed effects, which cause singleton observations to be dropped. Standard errors (in parentheses) are clustered at the startup level in all columns.

Overall, these tests broadly support our human RA scores, highlighting that our careful training and monitoring of the RAs resulted in effective data collection. However, given that the curves in Figure B.1 are not entirely overlapping, we further test the robustness of our findings. In particular, we replicated the same 2SLS analyses presented in Table VIII using measures of *Theory Score* and *Empirics Score* that are averaged between the three data sources (i.e., human RAs and two different state-of-the-art LLMs). Table B.II reports the results of these estimations, which largely confirm our previous findings, with the magnitudes of the estimated performance effects being larger but less precise.

APPENDIX C: ATTRITION TREATMENT

Attrition is a common challenge in RCTs, and our study addresses this issue using both descriptive and econometric approaches. Table C.I presents the overall response rates and attrition by treatment status. Given that *Termination* is a key outcome variable in our study, we add two rows to clarify the number of startups that cease their activities (i.e., they 'terminate') and the remaining 'net' attrition (i.e., participants that just stop providing data without confirming the fate of their startup). Moreover, the significant drop in the number of participants between the third and fourth rounds is mainly due to the India RCT having a shorter data collection time frame, due to specific logistical and organizational reasons.

TABLE C.I
Response rates and attrition numbers

Data collection round	0	1	2	3	4*	5*
# of responses (Sales variable)	1,556	1,182	979	782	597	520
% on surviving sample	100%	76%	83%	80%	84%	87%
# of terminations	-	44	61	71	54	12
% attrition (excl. terminations)	-	19%	11%	13%	9%	12%

Notes. This table briefly summarizes the response rates over time. *Please note that the response numbers for rounds 4 and 5 are negatively affected by the shorter data collection period for the India RCT, which stopped at round 3.

In addition, we perform selective attrition tests through a logit regression that relates attrition probabilities to treatment indicators and baseline covariates, excluding ones that are multicollinear with others. Estimates from Table C.II indicate that the odds of attrition – during the entire time frame of our RCTs – for the *TE* and *E* groups are, respectively, 2.06 and 2.07 times lower than for the control group ($exp(\beta_{TE}) = 2.06, exp(\beta_E) = 2.07, p = 0.000$), or approximately 17 percentage points less. This was to be expected, since the pure control group was – by design – not slated to receive any training while being subjected to the same data collection requests of the two treated groups.

TABLE C.II
Selective attrition tests

	Attrition
Control (reference)	–
	–
Theory-and-Evidence	-0.72
	(0.14)
Evidence	-0.73
	(0.14)
Cumulative Sales (Log €)	0.01
	(0.02)
Startup Employees	0.00
	(0.01)
Months of Operations	0.00
	(0.00)
Startup Sells a Product	0.02
	(0.14)
Startup Sells a Service	0.21
	(0.12)
Startup is For-Profit	0.04
	(0.20)
Perceived Prob. of Termination	-0.00
	(0.00)
Gender (=1 if Female)	0.11
	(0.11)
Bachelor's Degree or Higher	0.30
	(0.14)
Working Full Time	0.12
	(0.12)
Work Experience (Years)	0.00
	(0.01)
Development Stage (1-5)	-0.04
	(0.04)
Theory Score	0.00
	(0.06)
Empirics Score	0.01
	(0.05)
Constant	0.06
	(0.32)
Observations	1,554

Notes. This table reports logit selective attrition tests to estimate whether the probability of attrition systematically differs based on the treatment assignment or on other covariates. The reference treatment category is the Evidence-based group, and standard errors (in parentheses) are clustered at the startup level.

Despite receiving access to some generic materials on entrepreneurship and the promise of future access to exclusive events with expert guests, many startups in the control group eventually

stopped responding to our surveys and interviews. More importantly, Table C.II indicates that the two treatment groups have virtually identical attrition rates. This means that, conditional on being assigned to one of the two treatment groups, attrition rates were not different between the *TE* and *E* conditions during our study's time frame. This is crucial, and the lack of differential attrition rates for the two treated groups – central to our study's discourse – implies that there is no need to compute sharp 'Lee bounds' (Lee, 2009). Moreover, Table C.II indicates that the probability of attrition is not associated with the large number of covariates that we observe, increasing our confidence that attrition was mainly random. The only statistically significant association is with the education variable, indicating that the odds of attrition for founders with at least a Bachelor's degree are 35% higher than those without one $(exp(\beta) = 1.35, p = 0.039)$, or about 7 percentage points. This is also an unsurprising result, since entrepreneurship is a very risky career path and founders with higher education levels are more likely to have better outside options in terms of employment.

To assess the robustness of our findings, we also follow best practices in the literature on RCTs and employ inverse probability weighting (IPW, John, Gottschalk, and Moffitt, 1998) to account for attrition in our set of data. To preserve efficiency of the estimators and avoid variance inflation from extreme weights, we follow Hirano, Imbens, and Ridder, 2003 and estimate nonparametric propensity scores. In particular, we estimate propensity scores (i.e., probability of each observation being in our data set) from a nonparametric kernel regression using the only two statistically significant predictors of attrition, as identified in Table C.II (i.e., treatment assignment and education degree)[15]. We then estimate the ITT effects on performance employing weighted difference-in-differences regressions, where the sample weights are computed as the inverse of the propensity scores. Table C.III presents results that closely match the ones reported in Section 5. Following our previous models (and as in Equation 6), all the models in Table C.III include startup and time fixed effects, as well as standard errors clustered at the startup level.

[15]We also ran an alternative estimation of the propensity score adding the *"Startup Sells a Service"* variable which was barely significant at the 10% confidence level in Table C.II and results did not change.

TABLE C.III
Robustness tests with Inverse Probability Weights

	Cumulative Sales (Log €)	Employees	Log(Sales / Employees)	TFP
	(1)	(2)	(3)	(4)
TE × Post	0.53	0.21	0.62	0.52
	(0.16)	(0.24)	(0.18)	(0.16)
E × Post	0.05	0.22	0.11	0.12
	(0.14)	(0.26)	(0.16)	(0.15)
Constant	3.70	4.91	3.26	-0.10
	(0.04)	(0.06)	(0.04)	(0.04)
$\beta_{TE} = \beta_E$ (p-value)	0.002	0.969	0.004	0.011
Observations	5,270	5,104	5,104	5,104
Startup FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes

Notes. This table reports weighted difference-in-differences robustness analyses of the effect of the treatments on employment and productivity. We weight observations by the inverse of the propensity score, estimated with a nonparametric kernel regression. For Column (1), the revenue outcome is a measure of cumulative sales reported in logged Euros (€). For Column (2), the employment outcome is the count of the total number of people working for the startup at each data collection point. For Column (3), the productivity outcome is the log of the ratio between startup total sales (in Euro) and the total number of people working for the startup. For Column (4), the productivity outcome is a proxy estimation of Total Factor Productivity (TFP), computed as the residuals from a weighted OLS regression of Log Cumulative Sales on Log Employees (controlling for time and RCT location fixed effects). Each column also reports the two-sided p-value of Wald tests for the equality of the two treatment coefficients. All models include startup and time fixed effects, as well as standard errors (in parentheses) clustered at the startup level.

To further mitigate the risk that attrition and missing data pose to our inferences, we employ the Last-Observation Carried Forward (LOCF) technique – common in medical and clinical RCTs – whereby we fill missing observations with the last available data point for each attritor, resulting in a sample of $N = 8,946$. This is a simple imputation technique that allows us to perform a rapid robustness check of our results, as well as potentially incorporating information on the outside option value of entrepreneurs who terminate, as discussed in our model in Section 2.1. We then estimate our standard difference-in-differences regressions, following Equation 6. Estimates from Columns (1), (3) and (4) in Table C.IV confirm that our *TE* treatment's positive effect on revenue and productivity is still statistically significant across all models ($p = 0.000$), albeit with slightly

lower magnitudes. Interestingly, the coefficient estimates for the *E* treatment are now larger in magnitude and approach statistical significance at conventional levels ($p = 0.058$ in Column (1), $p = 0.034$ in Column (3), and $p = 0.059$ in Column (4)). Wald tests confirm that *TE* coefficients are still significantly larger than *E* ones ($p < 0.01$ in all three models). Consistently with our main results, none of the two treatments have any effect on startup employment levels, compared to the control group.

TABLE C.IV
Robustness tests with filled panel

	Cumulative Sales (Log €)	Employees	Log(Sales / Employees)	TFP
	(1)	(2)	(3)	(4)
TE × Post	0.41	-0.13	0.46	0.40
	(0.09)	(0.15)	(0.10)	(0.09)
E × Post	0.14	0.09	0.17	0.15
	(0.07)	(0.14)	(0.08)	(0.08)
Constant	3.25	4.78	2.81	-0.10
	(0.02)	(0.04)	(0.03)	(0.03)
$\beta_{TE} = \beta_E$ (p-value)	0.004	0.177	0.006	0.007
Observations	8,946	8,946	8,946	8,946
Startup FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
Estimation Panel	Filled	Filled	Filled	Filled

Notes. This table reports difference-in-differences robustness analyses of the effect of the treatments on employment and productivity. To mitigate the impact of attrition, we replace missing observations with the last available observation for each attritor (LOCF), resulting in a sample of 8,946 observations. For Column (1), the revenue outcome is a measure of cumulative sales reported in logged Euros (€). For Column (2), the employment outcome is the count of the total number of people working for the startup at each data collection point. For Column (3), the productivity outcome is the log of the ratio between startup total sales (in Euro) and the total number of people working for the startup. For Column (4), the productivity outcome is a proxy estimation of Total Factor Productivity (TFP), computed as the residuals from an OLS regression of Log Cumulative Sales on Log Employees (controlling for time and RCT location fixed effects). Each column also reports the two-sided p-value of Wald tests for the equality of the two treatment coefficients. All models include startup and time fixed effects, as well as standard errors (in parentheses) clustered at the startup level.

Given that LOCF is a relatively naïf imputation approach, we perform a further robustness

analysis using multiple imputation (MI, Little and Rubin, 2019; Imbens and Rubin, 2015). There is an ongoing debate in the literature regarding whether outcome variables should be imputed at all, due to their sensitivity to violations of the imputation assumptions. This is especially true when dropouts can be related to the treatment, which is partially the case in our experimental design (i.e., where *Termination* is a key outcome variable). After these words of caution and given that MI techniques are still a potentially relevant solution to the problem of attrition in longitudinal studies with drop-outs (Little and Yau, 1996), we also report results with MI estimations for completeness. Specifically, we employ a truncated regression model with 20 repetitions for the imputation process, which is well-suited to the characteristics of our key *Cumulative Sales (Log €)* dependent variable. To predict the values to impute, we use startup dummies, time dummies, and lagged sales values (because of the cumulative nature of our dependent variable).

TABLE C.V
Robustness tests with multiple imputations panel

	Cumulative Sales (Log €)	
	(1)	(2)
Theory-and-Evidence × Post	0.65	0.42
	(0.22)	(0.21)
Evidence × Post	0.35	–
	(0.21)	–
Constant	2.54	2.40
	(0.09)	(0.10)
$\beta_{TE} = \beta_E$ (p-value)	0.181	–
Observations	8,946	6,846
Startup FE	Yes	Yes
Time FE	Yes	Yes
Estimation Panel	Imputed, Full	Imputed, Treated

Notes. This table reports difference-in-differences robustness analyses of the effect of the treatments on employment and productivity. To mitigate the impact of attrition, we replace missing observations using a multiple imputation procedure consisting of 20 chained imputations. Imputed values are estimated using truncated regressions based on individual dummies, time dummies, and lagged sales values. In Column (1), the difference-in-differences model is estimated on the full sample with three treatment arms. In Column (2), we only restrict the sample to the two groups assigned to the *TE* or *E* treatments. All models include startup and time fixed effects, as well as standard errors (in parentheses) clustered at the startup level.

Column (1) of Table C.V estimates the same regression as Column (1) of Table V using the filled panel data set with 20 imputations. Results show that the *TE* treatment causes a 1.91 times higher sales growth rates compared to the control group ($exp(\beta_{TE}) = 1.91, p = .003$), which is even higher than the 1.71 times observed conditional on survival. Interestingly, also with this specification – following the trend with the LOCF imputation above – the *E* treatment's effect is larger in magnitude compared with Table V, albeit its statistical significance is still broadly insufficient ($p = 0.095$). Linear combination tests for the equality of the two treatment coefficients across the 20 imputations in Column (1) of Table C.V this time fail to reject the null hypothesis ($p = 0.181$). However, when estimating our difference-in-differences regression just on the sample of startups assigned to a treatment condition (i.e., dropping all the observations assigned to the control group), regression results in Column (2) of Table C.V confirm that the *TE* group's revenue growth rate is 1.52 times greater than that of the *E* group ($exp(\beta) = 1.52, p = 0.046$).

APPENDIX D: ROBUSTNESS OF INFERENCE AND MULTIPLE TESTING ADJUSTMENTS

In this Appendix, we introduce two complementary approaches to assess the robustness of our main findings. First, we employ Fisherian Randomization Inference to obtain exact finite-sample p-values that do not rely on asymptotic approximations or distributional assumptions. Second, we apply multiple hypothesis testing (MHT) adjustments to account for the multiplicity of outcomes and treatment comparisons in our experimental design, ensuring that our inference maintains appropriate control of error rates across families of tests.

D.1 Fisherian Randomization Inference (FRI)

To assess the robustness of our core findings about the effects of the *TE* and *E* treatments on startup revenue growth from Column (1) of Table V, we employ Fisherian Randomization Inference (Fisher, 1935) as a complementary approach to our main specifications. FRI provides exact finite-sample p-values by comparing our observed treatment effects to the distribution of all possible treatment assignments that could have occurred under our randomization protocols, testing the sharp null hypothesis that neither treatment affected any firm's revenue growth. This approach is particularly valuable for our analysis given the potential presence of outliers in revenue data, as FRI makes no distributional assumptions and provides valid inference regardless of sample size or the underlying data-generating process (Young, 2019). FRI ensures that our conclusions about the effectiveness of the *TE* and *E* treatments are not artifacts of asymptotic approximations or parametric assumptions, but rather reflect genuine treatment effects that are detectable through the randomization mechanism alone.

Table D.I reports the FRI tests of the key coefficients estimated in Table V. We perform 2,000 permutations as research shows "no appreciable change in rejection rates beyond 2,000 draws" (Young, 2019, p.572). Our results firmly reject the sharp null hypothesis of no *TE* treatment effect ($p = 0.010$) but fail to reject the null for the *E* treatment ($p = 0.858$), confirming our results from Table V. Figure D.1 also provides a visual representation of our FRI test permutations

through Kernel density plots, where the vertical red lines represent the parameter estimates in the actual sample.

TABLE D.I
FRI Test Results

Statistic	T_{obs}	c	n	$p = c/n$	$SE(p)$	Interval
β_{TE}	0.534	20	2,000	0.0100	0.0022	[0.0061, 0.0154]
β_{E}	0.055	1,716	2,000	0.8580	0.0078	[0.8419, 0.8730]

Notes. T_{obs} represents the observed test statistic (i.e., β coefficients from Column (1) of Table V), c is the count of $|T| \geq |T_{obs}|$, n is the number of permutations, $p = c/n$ is the p-value of the test, and $SE(p)$ is the standard error of the proportion. Confidence intervals for p are computed at the 95% level. This table was produced in Stata using the 'ritest' package from Heß, 2017.

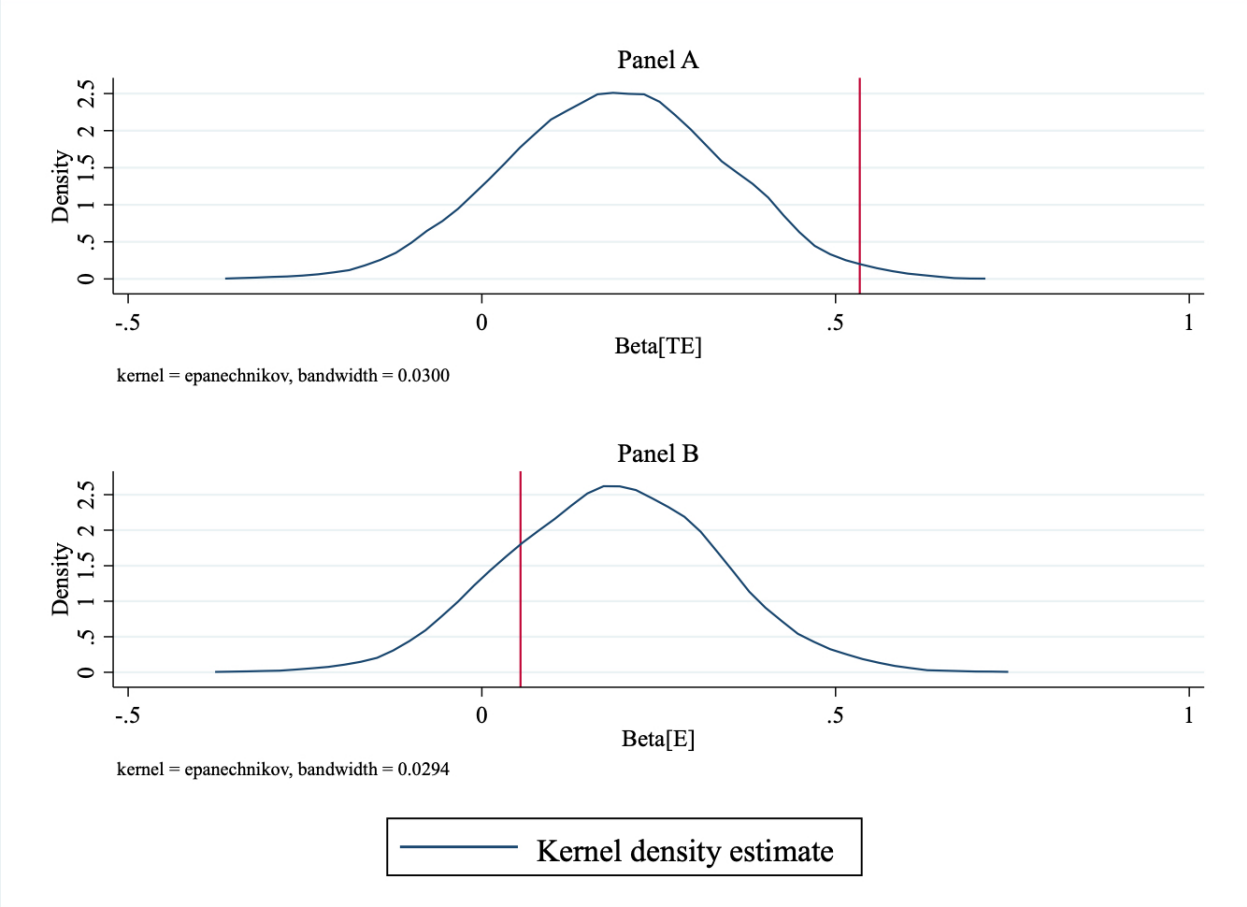


Figure D.1
Kernel Density Plots of FRI test permutations. The vertical red lines represent the parameter estimates in the actual sample.

D.2 Multiple Hypothesis Testing Adjustments

In this subsection, we adjust for multiple hypothesis testing (MHT) to mitigate the risk of inflating Type I error rates that can arise when simultaneously evaluating several outcomes – namely, revenue, employment, and productivity. By employing robustness adjustments, we enhance the credibility of our statistical inferences and affirm that the observed significant effects are indeed reflective of genuine causal impacts, in line with our theoretical framework and predictions.

The three-arm experimental design of our randomized controlled trials implies that there are nine hypotheses (with respective p-values) when testing our three key outcomes. For this reason, we start by following the procedure detailed by List, Shaikh, and Xu, 2019 to control for the familywise error rate (FWER). Their approach is particularly suitable for our context because it is designed for multiple outcomes and multiple treatments, although it does not allow for the inclusion of control variables nor fixed effects – unlike our regressions in Section 5. Table D.II reports the output of our adjustment procedure. Column (1) presents the between-groups difference-in-differences for all the dependent variables and possible treatment pairs. Column (2) then reports the unadjusted p-values, while Column (3) returns the multiplicity-adjusted p-values, calculated following Theorem 3.1 from List, Shaikh, and Xu, 2019. Despite the adjustments, all the p-values for the revenue and productivity outcomes in Column (3) remain highly statistically significant, confirming our key results. Moreover, adjusted employment p-values remain non-significant, as was the case in previous analyses.

Given that our core regressions included startup and time fixed effects (as described in Equation 6), we perform further MHT adjustments that allow for fixed effects. Specifically, we employ a Romano-Wolf stepdown adjustment, which is a sequential procedure that exploits the joint dependence among test statistics to control the family-wise error rate. It adaptively rejects null hypotheses in a stepwise manner, yielding enhanced power compared to traditional corrections when tests are correlated (Clarke, Romano, and Wolf, 2020). Given that this procedure does not allow for multiple treatments, unlike the one employed in Table D.II, we focus on the compari-

TABLE D.II
MHT Adjustments – familywise error rate (FWER)

Outcome	Comparison Groups	DI	p-values	
			Unadj.	Multiplicity Adj.
		(1)	(2)	(3)
Cumulative Sales (Log €)	TE vs Control	1.750	0.000	0.000
Cumulative Sales (Log €)	E vs Control	1.095	0.000	0.000
Cumulative Sales (Log €)	TE vs E	0.655	0.009	0.032
Employees	TE vs Control	0.653	0.044	0.107
Employees	E vs Control	0.642	0.088	0.144
Employees	TE vs E	0.011	0.982	0.982
Log(Sales/Employees)	TE vs Control	1.896	0.000	0.000
Log(Sales/Employees)	E vs Control	1.141	0.000	0.000
Log(Sales/Employees)	TE vs E	0.755	0.005	0.021

Notes. This table reports MHT tests based on pairwise tests. DI refers to difference-in-differences in means. This table was produced in Stata using the 'mhtexp' package from List, Shaikh, and Xu, 2019. The Multiplicity Adjusted p-values are computed using Theorem 3.1 from List, Shaikh, and Xu, 2019. Notably, this package does not closely reflect our core regressions as it does not allow for inclusion of fixed effects nor control variables. It therefore reports between-groups differences in means.

son between *TE* and *E* treated groups – which is at the core of our paper. Column (1) of Table D.III reports the unadjusted p-values of the difference-in-differences regressions comparing the *TE* and *E* treated groups, using startup and time fixed effects as well as standard errors clustered by startup. Column (2) then reports the Romano-Wolf adjusted p-values, while Column (3) additionally presents the p-values corresponding to the Holm multiple hypothesis correction. All the adjusted p-values for our revenue and productivity outcomes, although greatly inflated by the MHT adjustments, remain statistically significant, confirming that our results were not due to false rejections of the null hypotheses.

TABLE D.III

MHT Adjustments – Romano-Wolf stepdown

Outcome Variable	Unadjsted p-value	Romano-Wolf p-value	Holm p-value
	(1)	(2)	(3)
Cumulative Sales (Log €)	0.002	0.010	0.030
Employees	0.930	0.822	0.822
Log(Sales/Employees)	0.002	0.010	0.020

Notes. This table shows adjusted p-values computed in Stata using the 'rwolf' package from Clarke, Romano, and Wolf, 2020. Column (1) reports unadjusted p-values of our baseline difference-in-differences regressions with startup and time fixed effects, as well as standard errors clustered by startup. Column (2) then shows Romano-Wolf stepdown adjusted p-values, which control the FWER and allows for dependence among p-values by bootstrap resampling. Column (3) presents the p-values corresponding to the Holm multiple hypothesis correction.

APPENDIX E: EXPLORING TREATMENT HETEROGENEITY

Although not central to our paper's discourse and goals, we explore potential heterogeneity in the effect of our *TE* treatment. In fact, while the ITT results clearly indicate significant performance benefits of the *TE* treatment compared to the *E* one, it might have an uneven impact given that theory-based reasoning is not common practice among entrepreneurs and requires more cognitive effort. We employ the *"GenericML" R* package described and applied by Chernozhukov et al., 2025 to explore heterogeneity in our *TE* treatment's effect. This package involves rigorous and robust machine learning techniques to estimate key features of heterogeneous treatment effects in RCTs. In particular, it examines three measures relevant to both applied researchers and policy-makers: (i) the "Best Linear Predictor" (BLP), which estimates the conditional average treatment effect (CATE) using an ML-based proxy predictor; (ii) Sorted Group Average Treatment Effects (GATES), which capture average treatment effects across groups sorted by predicted heterogeneity; and (iii) Classification Analysis (CLAN), which describes the average characteristics of the most and least affected units. Overall, these measures allow us to determine whether treatment effect heterogeneity is detectable from observables, estimate treatment effects for different segments, and identify which covariates are most associated with this variation.

We train the models using a broad set of learners (i.e., random forest, lasso, elastic net, support vector machine, ranger random forest, extreme gradient boosting) with 250 splits of the data, as suggested by Chernozhukov et al., 2025. Consistent with our main analyses, we employ a difference-in-differences treatment estimator with treatment and time fixed effects, as well as standard errors clustered at the startup level. Differently from Equation 6, we did not include startup fixed effects as they would prevent studying heterogeneity across time-invariant covariates. Given that the GenericML package only supports a binary treatment, we focused once again on the subset of entrepreneurs assigned to the two treatment conditions as the *E* treated group is the more challenging and appropriate counterfactual.

When training the models through the GenericML specifications presented above, we inspect for heterogeneity along a vector Z_i comprising the following five covariates:

- Number of months that the startup has been operational for
- Gender of the main founder
- Years of total work experience of the main founder
- Dummy variable indicating whether the main founder has at least a college degree
- RCT location.

The GenericML routine presents us with several results. First, it estimates the following BLP regression:

$$Y_{ijt} = \beta_1 \times D_{ijt} + \beta_2 \times D_{ijt} \times Z_i + FE_{it} \tag{10}$$

where Y_{ijt} is the *Cumulative Sales (Log €)* outcome variable for startup i – assigned to treatment j – at time t, D_{ijt} is the classic difference-in-differences estimator that indicates whether the treatment j has affected the startup i at time t, and FE_{it} represents fixed effects for time and treatment. In this equation, β_1 represents the estimate of the average treatment effect (ATE). If the *TE* treatment's heterogeneity is minimal or absent, the effect of D_{ijt} is captured entirely by β_1, and the estimate of β_2 should approach zero. Conversely, if heterogeneity exists and the model successfully captures it using variables from the vector Z_i, β_2 should be positive. Thus, β_2 serves as a test for the presence of heterogeneity in our treatment's effects.

Panel A in Figure E.1 shows that the ATE is positive and statistically significant ($\beta_1 = 0.57, p = 0.029$), as well as closely aligned with the results presented in the other sections and analyses of this paper. Moreover, the estimate of β_2 is positive and different from zero ($p = 0.000$), indicating that there is significant heterogeneity in our *TE* treatment's effect along the Z_i vector of covariates. Given this evidence, we then leverage the estimates of Z_i to categorize observations based on their predicted treatment effects. To facilitate interpretation, we implement a median split, though results remain consistent when using quartiles. For observations with above- and below-median predicted effects, GenericML computes the GATES, allowing us to identify where treatment effects are most pronounced, absent, or negative. Panel B in Figure E.1 indicates that

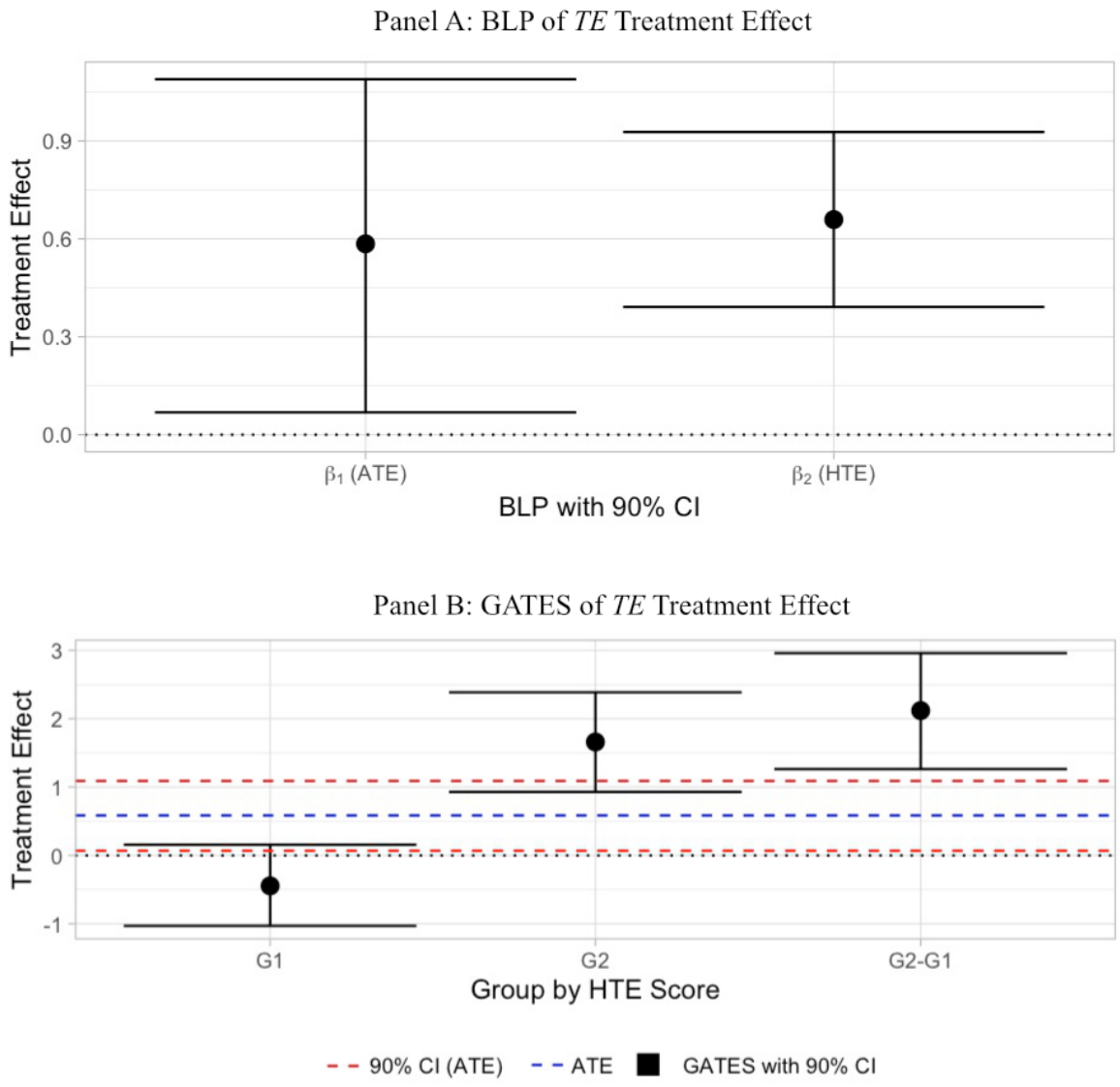


Figure E.1
This figure shows key plots compiled with the 'GenericML' R package. Panel A presents estimates of β_1 (ATE) and β_2 (HTE) from Equation 10. Panel B graphs the GATES using a median split across the distribution of treatment effects.

the GATE estimate for startups which have below-median treatment effects is negative, although not statistically different from zero ($G_1 = -0.46, p = 0.135$). Conversely, the GATE estimate for startups with above-median effects is positive and highly statistically significant ($G_2 = 1.65, p = 0.000$). The final estimate in Panel B shows that the difference between G_2 and G_1 is positive and significant ($p = 0.000$), confirming the presence of heterogeneity in the effect of our treatment.

To further investigate these differences, we compare the average covariate values between the two groups. This approach helps reveal how observed characteristics correlate with the heterogeneity detected through BLP and GATE estimates. Starting from the *Months of Operations* variable, the CLAN indicates that startups in the above-median treatment effect group are on average 3.15 months older than startups in the below-median group ($p = 0.001$). Although statistically significant, this difference is mostly negligible in terms of magnitude. CLAN results also signal a slightly higher proportion of females in the above-median group, equal to 3.9 percentage points or about 12% more ($p = 0.047$). The two groups do not differ in terms of years of work experience ($\Delta = 0.28, p = 0.430$), but there are on average 5% fewer founders with a college degree or higher among the above-median treatment effect group ($\Delta = -0.05, p = 0.002$). There might therefore be a potential partial substitution effect for those entrepreneurs who were already trained on causal reasoning in their university studies. The classification analysis (CLAN) also indicates the presence of heterogeneity with respect to our eight RCT locations ($p = 0.000$), with Colombia- and Italy-based startups exhibiting the highest average treatment effects, while Indian and Chinese ones show the lowest predicted average effects – reflecting potential heterogeneity in the quality of the local instructors.