THEORY-BASED REASONING AND PERFORMANCE[1]

ARNALDO CAMUFFO

ALFONSO GAMBARDELLA

DIEGO JANNACE

July 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 theoretical framework of decision-making under uncertainty, arguing 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 1,552 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 monthly 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Preliminary draft, please do not quote or circulate without authors' permission.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 provide a simple framework showing how a causal theory can affect entrepreneurial outcomes through two margins: it can improve selection by making posterior beliefs about venture quality more informative, and it can improve productivity by helping entrepreneurs allocate attention toward truly causal antecedents of performance. We then empirically test the propositions of our framework. Our study analyzes data from eight pre-registered randomized controlled trials (RCTs) involving 1,552 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 causal theories before testing them with 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 three-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

[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.

on an analysis of covariance (ANCOVA) approach with multiple post-intervention measurements (McKenzie, 2012) to estimate intention-to-treat effects on performance. We supplement these analyses with a battery of alternative specifications and robustness tests, including exact randomization inference, adjustments for attrition, and multiple-hypothesis testing. Our theoretical framework predicts that a theory-based approach generates both a selection and treatment effect, and provides the conditions to identify them empirically. We then find evidence of both effects. We also develop novel measures of theory-based causal reasoning and empirical testing, which we leverage 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. Conditional on survival, average monthly sales of the *TE* group are around 59% higher than the *C* group, driven by an increase in productivity rather than employment growth. This is equivalent to an average increase of about €970 per month during our data collection window, which is a meaningful effect in a sample where most ventures start from zero sales.

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.

THEORETICAL FRAMEWORK

2.1 Background

We model the implications of the interventions by distinguishing between two phases. In the first phase, early-stage entrepreneurs learn about the payoff from pursuing their venture. At each stage they can either commit to a decision – terminate the venture and take the outside option, or invest in the venture – or run another experiment and update their beliefs. In the second phase, entrepreneurs

who invest operate the venture and allocate their managerial time across activities.[2]

Let $g \in \{C, E, TE\}$ index the experimental group: control, evidence-based, and theory-and-evidence-based. Random assignment implies that the distribution of initial opportunities and founder characteristics is closely comparable across groups[3]. For notational simplicity, we write the initial prior mean as common across groups; all group differences below arise from the information and decision procedures induced by the interventions.

We normalize the outside option to zero. If the entrepreneur commits at stage i, the relevant object is the posterior expected present value of the venture, after optimizing all variable inputs other than the managerial choices modeled below. Denote this value by μ_i^g. The payoff from committing is therefore

$$v(\mu_i^g) = \max\{0, \mu_i^g\}, \tag{1}$$

where the entrepreneur terminates if $\mu_i^g \leq 0$ and invests if $\mu_i^g > 0$. This normalization avoids placing sign restrictions on a primitive productivity parameter: all fixed costs, optimized variable inputs, and continuation values are absorbed into the scalar payoff μ_i^g.

Entrepreneurs do not initially know the payoff from the venture. Let $\mathscr{I}_i^g$ denote the information available to an entrepreneur in group g at stage i, and let

$$\mu_i^g = \mathbb{E}[y \mid \mathscr{I}_i^g] \tag{2}$$

be the posterior mean of the payoff from committing to the venture. We write

$$y = \mu_i^g + \varepsilon_i^g, \qquad \mathbb{E}[\varepsilon_i^g \mid \mathscr{I}_i^g] = 0, \tag{3}$$

where ε_i^g is the residual uncertainty. Running an experiment at stage $i+1$ produces a signal that

[2]Some firms in our sample do not start from scratch, as implied by the fact that, at baseline, about one-third have positive sales. These small firms participated in the program because they wanted to introduce important changes. In the model, they learn about these changes in the first phase while generating profits in the old way; in the second phase, they either implement the changes and obtain the new payoff, or do not implement them and keep the old payoff.

[3]We later verify that this is the case through extensive tests of balance, reported in Table IV.

updates beliefs according to

$$\mu_{i+1}^g = \mu_i^g + \xi_{i+1}^g, \qquad \mathbb{E}[\xi_{i+1}^g \mid \mathscr{I}_i^g] = 0. \tag{4}$$

Thus posterior means are martingales. We assume that the residual uncertainty has finite variance and vanishes as entrepreneurs learn: $\mathbb{E}[(\varepsilon_i^g)^2 \mid \mathscr{I}_i^g] = (\sigma_i^g)^2 < \infty$ and $\sigma_i^g \to 0$ as $i \to \infty$. Equivalently, the sequence of experiments eventually makes the stochastic component of the payoff negligible.

2.2 Learning and commitment

At the end of stage i, entrepreneurs choose whether to commit or to run another experiment. Let $c_g > 0$ be the cost of an experiment for group g, and let $\beta \in (0,1)$ be the discount factor. The value function is

$$V_i^g(\mu_i^g) = \max\left\{v(\mu_i^g),\ -c_g + \beta\,\mathbb{E}\left[V_{i+1}^g(\mu_{i+1}^g) \mid \mathscr{I}_i^g\right]\right\}. \tag{5}$$

The state includes the residual distribution of ε_i^g; to keep notation light, we suppress it and write the value function as a function of the posterior mean.

Equation (5) captures the option value of experimentation. An experiment is valuable because it can move the entrepreneur across the commitment threshold: a venture that currently looks unprofitable may become worth pursuing, or a venture that currently looks profitable may be revealed as unattractive. Experimentation is costly and delays payoff realization, so entrepreneurs stop experimenting once the value of additional information is small.

Lemma 2.1. *For each group g, there exists a stage after which entrepreneurs commit to either termination or investment rather than continue experimenting.*

Proof. Fix a history at stage i. The value of any future sequence of costly and delayed experiments is bounded above by the value of observing the residual uncertainty immediately and for free. Thus

$$\mathbb{E}\left[V_{i+1}^g(\mu_{i+1}^g) \mid \mathscr{I}_i^g\right] \le \mathbb{E}\left[v(y) \mid \mathscr{I}_i^g\right] = \mathbb{E}\left[v(\mu_i^g + \varepsilon_i^g) \mid \mathscr{I}_i^g\right].$$

Since $v(\mu) = \max\{0, \mu\}$ is one-Lipschitz,

$$\mathbb{E}\left[v(\mu_i^g + \varepsilon_i^g) \mid \mathscr{I}_i^g\right] - v(\mu_i^g) \leq \mathbb{E}\left[|\varepsilon_i^g| \mid \mathscr{I}_i^g\right] \leq \sigma_i^g.$$

Therefore the value of continuing satisfies

$$\begin{aligned} &-c_g + \beta \mathbb{E}\left[V_{i+1}^g(\mu_{i+1}^g) \mid \mathscr{I}_i^g\right] - v(\mu_i^g) \\ &\quad \leq -c_g + \beta\left(v(\mu_i^g) + \sigma_i^g\right) - v(\mu_i^g) \\ &\quad = -c_g + \beta\sigma_i^g - (1-\beta)v(\mu_i^g) \leq -c_g + \beta\sigma_i^g. \end{aligned}$$

Because $\sigma_i^g \to 0$ and $c_g > 0$, there exists a stage after which $\beta\sigma_i^g < c_g$. At every such history, continuation is strictly worse than commitment. Hence the entrepreneur eventually commits to termination or investment. □

The next proposition characterizes the commitment regions. Let

$$C_i^g(\mu) \equiv -c_g + \beta\, \mathbb{E}\left[V_{i+1}^g(\mu + \xi_{i+1}^g) \mid \mathscr{I}_i^g\right] \tag{6}$$

be the value of continuing from posterior mean μ. We impose the standard single-crossing condition that the value of experimentation is highest for intermediate beliefs: when the venture looks clearly unattractive or clearly attractive, the expected gain from another signal is lower than when the entrepreneur is close to indifferent between termination and investment.

Proposition 2.2. *Suppose that $C_i^g(0) > 0$ and that the value of experimentation satisfies the single-crossing condition. Then there exist two cutoffs $\underline{\mu}_i^g < 0 < \overline{\mu}_i^g$ such that: (i) if $\mu_i^g < \underline{\mu}_i^g$, the entrepreneur terminates; (ii) if $\underline{\mu}_i^g < \mu_i^g < \overline{\mu}_i^g$, the entrepreneur experiments; and (iii) if $\mu_i^g > \overline{\mu}_i^g$, the entrepreneur invests.*

Proof. As $\mu \to -\infty$, even perfect revelation of the remaining uncertainty has value converging to zero. Thus $C_i^g(\mu) < 0 = v(\mu)$ for sufficiently low μ, and the entrepreneur terminates. At $\mu = 0$, the assumption $C_i^g(0) > 0 = v(0)$ implies that experimentation is preferred. As $\mu \to +\infty$, the

continuation payoff grows at rate at most $\beta\mu$, while the payoff from investing immediately grows at rate μ. Since $\beta < 1$ and $c_g > 0$, investment is preferred for sufficiently high μ. The single-crossing condition then yields the two cutoffs. □

The proposition implies the usual experimentation logic. Entrepreneurs with very pessimistic beliefs terminate, entrepreneurs with very optimistic beliefs invest, and entrepreneurs with intermediate beliefs experiment. New signals move posterior beliefs upward or downward. If an updated posterior falls outside the experimentation region, the entrepreneur commits (either to termination or investment); otherwise, experimentation continues.

2.3 What the interventions change

The three experimental groups differ in their information technology, not in preferences. Entrepreneurs in the control group receive no structured training in experimentation or causal theory-building. We capture this as a higher cost of experimentation or less informative signals. Entrepreneurs in the evidence-based group learn to run experiments, reducing the cost of acquiring signals and increasing their informativeness. Entrepreneurs in the theory-and-evidence-based group receive the same experimental tools as the evidence-based group, and in addition learn to formulate causal theories before testing them.

Formally, the *E* intervention improves the signal technology relative to the *C* group: it weakly lowers c_g or increases the precision of ξ^g_{i+1}. The *TE* intervention adds a causal-theory signal T_i to the information available under evidence-based experimentation,

$$\mathscr{I}_i^{TE} = \mathscr{I}_i^{E} \vee \sigma(T_i). \tag{7}$$

The signal T_i represents a structured causal model linking antecedents and contextual factors to venture performance. We assume that T_i is informative in the Blackwell sense: an entrepreneur with $\mathscr{I}_i^{TE}$ can always ignore the theory and imitate an entrepreneur with $\mathscr{I}_i^{E}$, but can sometimes make strictly better decisions by using the additional causal structure. This assumption does not

require theories to be correct with probability one. It requires that, in expectation, the theory-building procedure taught in the intervention produces decision-relevant information rather than pure noise.

2.4 Performance

We now unpack the operating payoff generated after investment. Entrepreneurs have one unit of managerial time to allocate across m possible activities. Let $\theta^* = (\theta_1^*, \ldots, \theta_m^*)$ denote the true causal productivity of these activities, with $\theta_j^* \geq 0$. The component of the venture payoff affected by managerial time is

$$\sum_{j=1}^{m} \theta_j^* t_j^{\nu}, \qquad \nu \in (0,1), \qquad \sum_{j=1}^{m} t_j = 1. \tag{8}$$

Other components of venture quality are included in the scalar posterior mean described above. The role of causal theory is to help entrepreneurs learn which activities have high causal productivity and which do not.

At stage i, an entrepreneur in group g forms posterior beliefs

$$m_{ij}^g = \mathbb{E}[\theta_j^* \mid \mathscr{I}_i^g]. \tag{9}$$

If the entrepreneur were to operate the venture at this stage, the optimal allocation of time solves

$$\max_{t_1,\ldots,t_m} \sum_{j=1}^{m} m_{ij}^g t_j^{\nu} \quad \text{subject to} \quad \sum_{j=1}^{m} t_j = 1. \tag{10}$$

The solution is

$$t_{ij}^{g*} = \frac{(m_{ij}^g)^{1/(1-\nu)}}{\sum_{k=1}^{m} (m_{ik}^g)^{1/(1-\nu)}} \tag{11}$$

for activities with positive posterior productivity, and zero for activities with $m_{ij}^g = 0$. The associ-

ated expected operating payoff from the time-allocation component is

$$\phi(m_i^g) = \left(\sum_{j=1}^{m} (m_{ij}^g)^{1/(1-\nu)} \right)^{1-\nu}. \tag{12}$$

The following theorem states the two channels through which causal theories affect performance.

Theorem 2.3. *Suppose that the causal-theory signal available to TE entrepreneurs is Blackwell more informative than the information available to E entrepreneurs about both the scalar venture payoff and the causal productivity vector θ^*. Then, theory-based reasoning weakly raises expected performance relative to evidence-based experimentation. The effect operates through two channels: (i) a selection effect, because more informative posterior beliefs improve termination and investment decisions; and (ii) a treatment effect, because more informative beliefs about θ^* improve the allocation of managerial time among activities. The inequalities are strict when the additional information is decision-relevant with positive probability.*

Proof. First consider the selection effect. Because $\mathscr{I}_i^{TE}$ is Blackwell more informative than $\mathscr{I}_i^E$, a TE entrepreneur can always ignore the additional theory signal and implement the same termination, experimentation, or investment rule available to an E entrepreneur. Therefore, for the same preferences, costs, and feasible actions, the expected value of the optimal commitment policy under $\mathscr{I}_i^{TE}$ is weakly higher than under $\mathscr{I}_i^E$. Equivalently, since

$$\mathbb{E}[\mu_i^{TE} \mid \mathscr{I}_i^E] = \mu_i^E,$$

the more informative posterior is a mean-preserving spread of the less informative posterior, and the convexity of the terminal payoff $v(\mu) = \max\{0, \mu\}$ gives

$$\mathbb{E}[v(\mu_i^{TE}) \mid \mathscr{I}_i^E] \geq v(\mu_i^E).$$

Thus more informative beliefs weakly improve the termination and investment decision. The im-

provement is strict when the additional information changes the optimal extensive-margin decision with positive probability.

Now consider the treatment effect, holding fixed the set of operating ventures. Let

$$m_i^{TE} = \mathbb{E}[\theta^* \mid \mathscr{I}_i^{TE}], \qquad m_i^{E} = \mathbb{E}[\theta^* \mid \mathscr{I}_i^{E}].$$

By the law of iterated expectations,

$$\mathbb{E}[m_i^{TE} \mid \mathscr{I}_i^{E}] = m_i^{E}.$$

The operating value

$$\phi(m) = \max_{\sum_j t_j = 1,\, t_j \geq 0} \sum_{j=1}^{m} m_j t_j^{\nu}$$

is convex in m, because it is the pointwise maximum of functions that are linear in m. Therefore Jensen's inequality implies

$$\mathbb{E}[\phi(m_i^{TE}) \mid \mathscr{I}_i^{E}] \geq \phi(m_i^{E}).$$

Taking expectations yields

$$\mathbb{E}[\phi(m_i^{TE})] \geq \mathbb{E}[\phi(m_i^{E})].$$

The inequality is strict whenever no single time allocation is optimal for all posterior productivity vectors generated by the additional theory signal with positive probability. □

The theorem highlights the economic relevance of theory-based reasoning. Causal theories are useful not because they mechanically make entrepreneurs more confident, but because they provide decision-relevant information about venture quality and about the activities that causally affect performance. When causal antecedents are unevenly productive, this information will often lead entrepreneurs to place more weight on some activities and less on others. The performance gain, however, comes from the informativeness of this reallocation, not from concentration of attention per se.

2.5 Empirical implications

The model generates distinct predictions for the three experimental groups. First, relative to the control group, both treatment groups should have a higher value of experimentation because they face lower costs of acquiring and interpreting evidence or obtain more informative signals. In the notation above, this corresponds to a lower c_g, a more precise signal ξ^g_{i+1}, or both. This makes treated entrepreneurs less likely to terminate immediately when beliefs are intermediate, because experimentation preserves option value.

Second, as learning proceeds, experimentation can increase termination among ventures whose posterior beliefs deteriorate. The model therefore does not impose a monotone prediction that TE entrepreneurs must always terminate more or less often than E entrepreneurs. A causal theory has two countervailing effects on termination. It may reveal downside risk and induce earlier exit of weak ventures. But it may also identify high-return actions that increase the expected value of continuing. The sign of the overall termination effect is therefore an empirical object.

Third, the framework separates selection and treatment conceptually. The selection channel changes which ventures operate: more informative beliefs move low-posterior ventures toward exit and high-posterior ventures toward investment. The treatment channel changes how operating ventures perform: more informative beliefs about θ^* improve the allocation of managerial time and therefore raise productivity.

The following proposition summarizes the main empirical restrictions.

Proposition 2.4. *The model implies: (i) E and TE entrepreneurs have a higher value of experimentation than C entrepreneurs; (ii) relative to E, TE entrepreneurs make more informative termination and investment decisions, but the sign of the raw termination-rate difference is ambiguous; and (iii) any TE advantage in productivity among comparable operating ventures is evidence of the treatment channel rather than the selection channel.*

Proof. Part (i) follows directly from (5): lower experimentation costs or more informative signals raise the continuation value. Part (ii) follows from the value-of-information result in Theorem 2.3:

more informative beliefs improve decisions, but can either increase or decrease observed termination depending on whether the dominant new information concerns downside risk or upside opportunities. Part (iii) follows because selection changes the identity of operating ventures, whereas the treatment channel changes the optimal allocation in (11) and the resulting operating payoff in (12). Holding the operating venture and its posterior quality fixed, selection has no additional effect on productivity; only improved allocation of activities does. □

DATA SOURCES AND EXPERIMENTAL SETTING

The data for this paper were collected through eight field RCTs involving a total of 1,552 early-stage startups[4] 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 I). 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.

[4]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.

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 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 stratified randomization by RCT location 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) [5].

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 36 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 on 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'

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

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 three-month training program, while those in the control group received only informational content, such as news, videos, and webinars, without any formal training[6]. The interventions consisted of six half-day, in-person 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 II 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,

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

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 three-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 incentive 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 1,552 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 (12.4% of the sample), tourism (9.7%), and social media (8.0%). 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 codebook 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) *Monthly Sales*, which reports the average amount of monthly revenue in Euro[7] earned by each startup during any data collection period; (ii) *Employees*, which measures the number of founders, full-time, and part-time employees working for the startup; (iii) *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[8]. 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.

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*

[7]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.

[8]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.

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, 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 two of them returned a statistically significant difference at the typical 95% confidence level: at baseline, *E* startups are on average 2.78 months older than *TE* ones ($p = 0.03$), and they declare 3.56 more hours worked per week than control startups ($p = 0.04$). Such minor imbalance is to be expected for such a large number of pairwise tests, and the magnitude of the difference is not concerning. We also perform F-tests for joint significance across all balance variables for each test pair, which reassuringly 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 stratified by location.

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 periodic attrition rate in the five post-baseline data collection periods (conditional on survival) was 13.4%, ranging from 21% 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 1,552 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. Finally, we use measures of theory and experimentation quality to explore our theoretical mechanisms.

5.1 Performance

Our theorem postulates that causal theories lead *TE* startups to weakly outperform the other groups due to a combination of a selection and a treatment effect, while experiments should weakly raise performance for *E* startups compared to control ones through a selection mechanism. We empirically test these predictions via an ANCOVA approach, which is especially appropriate for RCTs with a baseline and several post-treatment measurements, as well as for outcomes such as business revenue or profits (McKenzie, 2012). Our main specification is the following pooled OLS regression:

$$Y_{it} = \beta_1 D_i^{\mathrm{TE}} + \beta_2 D_i^{\mathrm{E}} + \gamma Y_{iB} + \lambda_{rt} + \varepsilon_{it} \qquad \text{for } t > 1 \tag{13}$$

where Y_{it} is the performance outcome of interest for startup i at period t; D_i^{TE} and D_i^{E} are indicators equal to one if startup i is assigned to the *TE* or *E* treatment arm, respectively, with the control group (*C*) as the omitted reference category; Y_{iB} is the baseline value of the outcome, included as the ANCOVA covariate[9]; λ_{rt} are RCT-by-period fixed effects, with r indexing the RCT stratum to which startup i belongs and t the period; and ε_{it} is an error term, with standard errors clustered at the startup level. The coefficients of interest are β_1 and β_2, which measure the average effect of each treatment arm on performance relative to the control group. Conditioning on the baseline outcome Y_{iB} controls for persistent differences in startup performance and, by absorbing a large share of the residual variation, improves the precision of the estimated treatment effects; the RCT-by-period fixed effects λ_{rt} further absorb any stratum-specific shocks common to a given period. We estimate equation (13) separately for each performance measure.

5.1.1 Revenue

We begin by estimating the ITT effect of our treatments on startups' monthly sales (i.e., the average amount of monthly revenue in Euro earned by each startup during any given data collection period), a key performance parameter for early-stage ventures.

Given that the distribution of our revenue data presents a significantly skewed right tail (as apparent also from Table III), we start by estimating the treatments' effect on $log(1+MonthlySales)$, which accommodates observations equal to zero. Column (1) of Table V reports our OLS ANCOVA estimates, using the full panel of post-treatment observations. Conditional on survival[10] and pooling all post-treatment periods, the ANCOVA estimate in column (1) indicates that the *TE* treatment increased monthly sales by 63 log points – approximately 87% – relative to the control group, conditional on baseline sales ($\beta_{TE} = 0.63, \exp(\beta) = 1.87, p = 0.003$). The effect of the *E* treatment is instead directionally negative and not statistically distinguishable from zero ($\beta_E = -0.06, \exp(\beta) = 0.95, p = 0.793$).

[9] When using *Monthly Sales* in levels as the outcome variable for Poisson regressions, the ANCOVA covariate is composed of the log of baseline sales (when they are positive) and a dummy that indicates zero baseline sales.

[10] Appendix C then extensively discusses attrition and presents a battery of robustness analyses with different imputation techniques for missing data.

While log transformations are well-suited for skewed revenue data and offer ease of interpretation, the presence of zeros in the outcome data in levels makes the estimate sensitive to arbitrary rescaling of the units of Y, especially when the rescaling factor is large (Chen and Roth, 2024). While we do not rescale sales and simply report them in Euro units, Chen and Roth (2024) show that $log(1+y)$ estimates are still "not well-defined for individuals for whom the treatment increases their outcome from zero to a positive value" (p.892). Therefore, we follow their recommendation and also estimate the average treatment effect (ATE) in levels expressed as a percentage of the control mean via Poisson regression (Wooldridge, 2010). This specification is insensitive to the units of Y and handles zeros naturally, however it imposes no diminishing returns and it risks being overly influenced by startups in the right tail of a skewed distribution (Chen and Roth, 2024).

Columns (2) through (4) of Table V report Poisson ANCOVA estimates of the effects of the treatments on startups' *Monthly Sales*, in levels. The dependent variable in column (2) is the unwinsorized monthly sales, and estimates indicate that average sales of the *TE*-treated group are 45% higher than the control group, although this effect is imprecisely estimated due to extreme values in the right-tail of the distribution inflating the standard errors ($\beta_{TE} = 0.37, p = 0.213$). Estimates for the *E* treatment are once again directionally negative ($\beta_E = -0.48, p = 0.058$) and likely also influenced by the outliers.

To mitigate the outsize influence of right-tail outliers on our estimates, the outcome in column (3) is the level of monthly sales winsorized – by RCT-period to better address true outliers – at the 95th percentile only, as a mass point at zero makes left-tail winsorization uninformative. Poisson estimates from column (3) show that average monthly sales of the *TE* group are 59% higher than the *C* group ($\beta_{TE} = 0.47, p = 0.018$), conditional on survival. Sales of the *E* group are once again not statistically distinguishable from control ones ($\beta_E = 0.11, p = 0.640$). Translating the proportional Poisson estimate into euros, the average marginal effect of *TE* treatment assignment corresponds to an increase of approximately €970 in monthly sales per firm ($p = 0.012$).

When further winsorizing at the 90th percentile, estimates from column (4) show that the *TE* treatment's effect grows in magnitude ($\beta_{TE} = 0.51, p = 0.004$), while the *E* treatment's one

remains not distinguishable from zero ($\beta_E = 0.20, p = 0.315$). This indicates that our *TE* treatment is most effective for smaller and early-stage startups compared to more mature ones, consistent with the fact that the interventions' contents were tailored specifically for nascent ventures.

Given the early-stage nature of startups in our study, with 69% of them having zero baseline sales, the outcome in column (5) of Table V is a binary measure indicating whether a startup generated positive sales during our study's time frame. Linear probability results[11] show that the *TE* treatment increases the probability of a sale by seven percentage points on average ($\beta_{TE} = 0.07, p = 0.013$), or about 55% on the control group mean. 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). Even in this specification, the *E* treatment's effect is modest and indistinguishable from zero ($\beta_E = 0.01, p = 0.741$).

When comparing the two treatment coefficients across the different specifications of Table V, Wald tests of equality broadly reject the null hypothesis that they are equal. This indicates that the monthly sales of startups in the theory-treated group are higher than those of both the control and the evidence-based groups. 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 III presents the cumulative distribution functions (CDFs henceforth) of the logged *Monthly Sales* outcome variable for all startups in our study. Panel A plots the CDFs for the baseline period, while Panel B plots the CDFs for the post-treatment periods, consistent with our ANCOVA regressions in Table V. Panel A of Figure III shows that the revenue distribution of the two treated groups is virtually identical at baseline (two-sample Kolmogorov–Smirnov test results:

[11] Results are identical when using a Logit specification.

$D = 0.03, p = 0.989$). 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 to the right, compared to the *E* treatment (two-sample Kolmogorov–Smirnov test results: $D = 0.11, p = 0.002$).

To further investigate the time dynamics of the differential sales growth rates between the two treatment groups, we modify equation (13) by interacting the intervention and time period indicators, estimating the model on the subset of startups assigned to the *TE* and *E* groups. In this way, we present a more granular analysis of the treatment effect over time. Figure IV plots the coefficients of this regression (with 95% confidence intervals), where the blue curve represents the *TE* treatment and the counterfactual (i.e., the red line in $y = 0$) is the *E* treatment. This graph shows that the average sales of the *TE* group, conditional on survival, steadily increase over time compared to the ones of the *E* group. The differences begin to be statistically significant already from the first post-treatment period, approximately four months after the start of the training.

These results are consistent with our theoretical framework's empirical predictions, confirming that theory-based reasoning raises entrepreneurial performance compared to both control and evidence-treated entrepreneurs. Results in Table V also show that experimentation alone does not provide material benefits to entrepreneurs, potentially indicating that the experiments' signals are not informative enough (in terms of c_g and/or ξ^g_{i+1} enhancements) to generate a substantial performance improvement without the guidance of a causal theory. Empirically, the popularity of A/B tests and other experimentation techniques in recent entrepreneurship practice and online resources might also lead some control entrepreneurs to be 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, following our theoretical framework. Table VI presents the results of ANCOVA ITT regressions estimating

employment and productivity outcomes, adopting the same specification described in Equation (13).

Column (1) of Table VI shows that both *TE* and *E* treatments do not increase the number of startup employees compared to control firms ($\beta_{TE} = 0.10, p = 0.789; \beta_E = 0.53, p = 0.197$). 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. Furthermore, there are no significant differences across the experimental conditions with respect to the number of hours worked per week. This is consistent with the interpretation that the positive performance effects of the *TE* treatment displayed in Table V could be due to an increase in productivity, in line with our theorem.

Column (2) indeed shows that the productivity of *TE* startups (measured in Euro of revenue per employee) is 75% higher on average than that of *C* startups ($\beta_{TE} = 0.56, p = 0.005$). The ITT effect of the *E* treatment on startups' productivity is instead negative but not statistically significant ($\beta_E = -0.12, p = 0.533$). Wald tests of equality also confirm that the difference between the two treatment coefficients is statistically significant ($p = 0.000$). We then calculate a residual proxy measure of total factor productivity (TFP) by regressing logged sales on the log number of employees, with RCT $\times$ period 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 ANCOVA regression (as in Equation (13), 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 70% increase in average productivity compared to the control group ($\beta_{TE} = 0.53, p = 0.000$). The *E* treatment's effect is again mildly negative but not statistically significant ($\beta_E = -0.14, p = 0.301$). As above, the Wald tests reject the null hypothesis that the coefficients of the two treatments are equal ($p = 0.000$). 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, on average has no impact on productivity, mir-

roring previous results on revenue generation. 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 further verify robustness, Appendix C inspects attrition and tests countermeasures, Appendix D then 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

Given that the predictions of our theoretical framework on termination rates are ambiguous due to competing effects, rigorous empirical evidence is especially valuable. Figure V depicts 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 external 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 V 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 V are consistent with the reasoning behind our theoretical framework. *C* entrepreneurs are not trained to experiment or to formulate causal theories. They therefore face a lower value of experimentation, either because experiments are more costly, because signals are less precise, or because they do not recognize which signals are

decision-relevant. Thus, entrepreneurs in the control group are more likely to terminate at the outset, when they do not yet perceive the option value of experimentation, and less likely to terminate later, when treated entrepreneurs have accumulated more decision-relevant information about downside risk. Moreover, the *TE* and *E* curves in Figure V are not statistically different from each other ($\chi^2 = 0.00, p = 0.964$).

To supplement the Kaplan-Meier analyses in Figure V, 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 average probability of termination, compared to the control group ($p \approx 0.03$ for both coefficients). Logit average marginal effects in column (2) are virtually identical to the OLS ones reported in the first column. To account for the right-censoring of our survival data, we then estimate the time to termination through a Tobit model. Estimates in column (3) show that treated startups, on average, 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 V that outweigh the initial terminations by the control group. Overall, these estimates indicate that, for sufficiently ample time horizons, *TE* and *E* entrepreneurs are similarly likely to terminate their activities and more likely than *C* ones.

These patterns are consistent with our framework on two margins. First, the elevated termination rates in both treated arms relative to *C* are consistent with the selection channel: more informative posteriors – whether produced by experimentation alone or in combination with causal theory – move ventures with deteriorating beliefs toward exit. Second, the absence of a differential termination pattern between TE and E is accommodated by the ambiguity the model attaches to the sign of the additional termination effect of causal reasoning. Causal theory may reveal downside risk and accelerate the exit of weak ventures, but it may equally surface high-return actions that raise the expected value of continuing. The empirical equivalence of TE and E termination rates is consistent with these forces approximately offsetting on the selection margin. Under this

interpretation, the incremental returns to causal theory found in the previous section should operate through the treatment channel (i.e., productivity conditional on continuation) rather than through differential selection on exit.

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:

$$Theory\ Score_{it} = TE_i \times Period_t + S_i + I_t + \varepsilon_{it} \tag{14}$$

$$Empirics\ Score_{it} = TE_i \times Period_t + S_i + I_t + \varepsilon_{it} \tag{15}$$

where we interact the *TE* treatment indicator with the data collection period. S_i and I_t are the fixed effects for startup and interview period, respectively, and ε_{it} 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 VI 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 examine whether our theory and empirics indexes are empirically associated with performance outcomes. In Table VIII, we estimate the correlation between key startup outcomes – in terms of sales, employment, and productivity – and our *Theory* and *Empirics Score* measures. We do so by estimating equation (13), with the additional inclusion of controls for the baseline levels of our theory and empirics score variables. ANCOVA results in column (1) of Table VIII suggest that a one-unit increase in *Theory Score* is associated with a 7.8% increase of startup monthly sales ($\beta_{Theory} = 0.08, p = 0.019$). On the other hand, the *Empirics Score* measure does not appear to be correlated with our monthly sales outcome when keeping the *Theory Score* constant ($\beta_{Empirics} = 0.00, p = 0.944$). 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 ANCOVA model for employment and productivity outcomes. Estimates from column (2) then indicate that our 'theory' index is not correlated with startup employment ($\beta_{Theory} = 0.03, p = 0.693$), whereas our 'empirics' one is mildly associated with the total number of people working for the startups ($\beta_{Empirics} = 0.12, p = 0.059$). 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 monthly sales and the total number of employees. A one-unit increase in *Theory Score* is associated with a 6.5% increase

in the amount of sales per startup employee ($\beta_{Theory} = 0.06, p = 0.035$). Similarly to column (1), the estimated *Empirics Score* coefficient is not distinguishable from zero ($\beta_{Empirics} = -0.01, p = 0.635$). Productivity results from column (4) also show a positive correlation between the residual proxy for TFP and our 'theory' measure ($\beta_{Theory} = 0.06, p = 0.013$), but not our 'empirics' one ($\beta_{Empirics} = -0.02, p = 0.329$).

These exploratory results shed additional light on the theoretical mechanism behind the performance advantages of the *TE* treatment. The positive association between *Theory Score* and performance suggests that the *TE* treatment operates through a direct effect on entrepreneurs' decision-making capabilities: startups whose founders developed higher-quality causal theories achieved significantly higher revenue and productivity. This is consistent with a treatment effect channel – in which entrepreneurs learn to reason more effectively – rather than a selection channel in which the *TE* treatment merely induces differential exit. The absence of a positive association between *Empirics Score* and performance further sharpens this interpretation: it is the quality of causal reasoning, not experimentation, that drives the performance gains we observe. This is consistent with our model, in which experimentation entails costs c – including managerial time, monetary outlays for MVPs and large-scale tests, and potential revenue delays – that may offset its benefits when testing is not guided by a well-formulated theory.

CONCLUSION AND POLICY IMPLICATIONS

This study provides experimental evidence that theory-based decision making under uncertainty enhances firm performance. Across eight field experiments with 1,552 early-stage startups spanning diverse economies and sectors, entrepreneurs trained to formulate and test causal theories achieve superior outcomes in monthly sales and productivity, relative to those relying solely on iterative experimentation or operating without structured intervention. The capacity to theorize about cause and effect – not merely to collect data – emerges as an important determinant of economic success, consistent with the view that causal, theory-based reasoning is a cornerstone of effective decision making in uncertain environments (Gilboa, Minardi, and Samuelson, 2020).

We complement these estimates with a transparent back-of-the-envelope cost–benefit assessment. Using the ingredients-costing method of Dhaliwal et al. (2013) – valuing donated inputs at market prices and excluding all research expenditures – we estimate that replicating the program would cost approximately €286 per assigned startup (Appendix F). Against the estimated intention-to-treat effect of €970 in additional monthly sales, our central assumption of a 15 percent marginal profit rate implies €1,701 in discounted incremental profits per assigned startup within the twelve-month window we observe – roughly six times the program's cost, which is recouped within two months. Even under our most conservative assumption of a 5 percent marginal profit rate, 12-month benefits reach €567 per assigned startup, twice the program's cost, with no persistence of effects assumed beyond the observed data. The comparison between our treatment arms sharpens this implication: because the *TE* and *E* interventions were identical in duration, instructors, and delivery, the marginal cost of embedding causal-theory content in existing entrepreneurship training reduces to a one-time adaptation of curriculum and instructor training (€8 per assigned startup at our scale), while its marginal benefit is the full *TE–E* differential. These figures capture private returns and assume no displacement of untreated competitors, an assumption supported by market-level experimental evidence (McKenzie and Puerto, 2021); Appendix F details all assumptions and sensitivity analyses.

Several avenues merit further exploration. Extending the analysis to more mature firms and to additional contexts would help ascertain the generalizability of our results beyond the eight countries studied here. Integrating longer-term performance metrics and more granular evaluations of decision-making processes could illuminate the mechanisms through which theory-based reasoning translates into sustained competitive advantage, and combining our approach with complementary frameworks – behavioral or network-based – could clarify how causal reasoning interacts with other determinants of firm performance.

In sum, our results indicate that coupling causal theories with experimentation improves decision making under uncertainty, and that teaching entrepreneurs to do so passes a cost–benefit test under conservative assumptions.

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.

Alfonsi, Livia, Oriana Bandiera, Vittorio Bassi, Robin Burgess, Imran Rasul, Munshi Sulaiman, and Anna Vitali (2020), "Tackling youth unemployment: Evidence from a labor market experiment in Uganda", *Econometrica* 88.(6) 2020, 2369–2414.

Anderson, Stephen J and David McKenzie (2022), "Improving business practices and the boundary of the entrepreneur: A randomized experiment comparing training, consulting, insourcing, and outsourcing", *Journal of Political Economy* 130.(1) 2022, 157–209.

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.

Chen, Jiafeng and Jonathan Roth (2024), "Logs with zeros? Some problems and solutions", *The Quarterly Journal of Economics* 139.(2) 2024, 891–936.

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.

Dhaliwal, Iqbal, Esther Duflo, Rachel Glennerster, and Caitlin Tulloch (2013), "Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education", *Education policy in developing countries* 17 2013, 285–338.

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

Fitzgerald, John, 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.

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.

Gagliardi, Luisa and Elena Novelli (2025), "Female entrepreneurs targeting women: Strategic redirection under scientific decision-making", *Organization science* 36.(2) 2025, 718–736.

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.

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.

Horowitz, Joel L and Charles F Manski (2000), "Nonparametric analysis of randomized experiments with missing covariate and outcome data", *Journal of the American statistical Association* 95.(449) 2000, 77–84.

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).

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.

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.

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

McKenzie, David and Susana Puerto (2021), "Growing markets through business training for female entrepreneurs: A market-level randomized experiment in Kenya", *American Economic Journal: Applied Economics* 13.(2) 2021, 297–332.

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).

Wooldridge, Jeffrey M (2010), *Econometric analysis of cross section and panel data*, (MIT press).

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

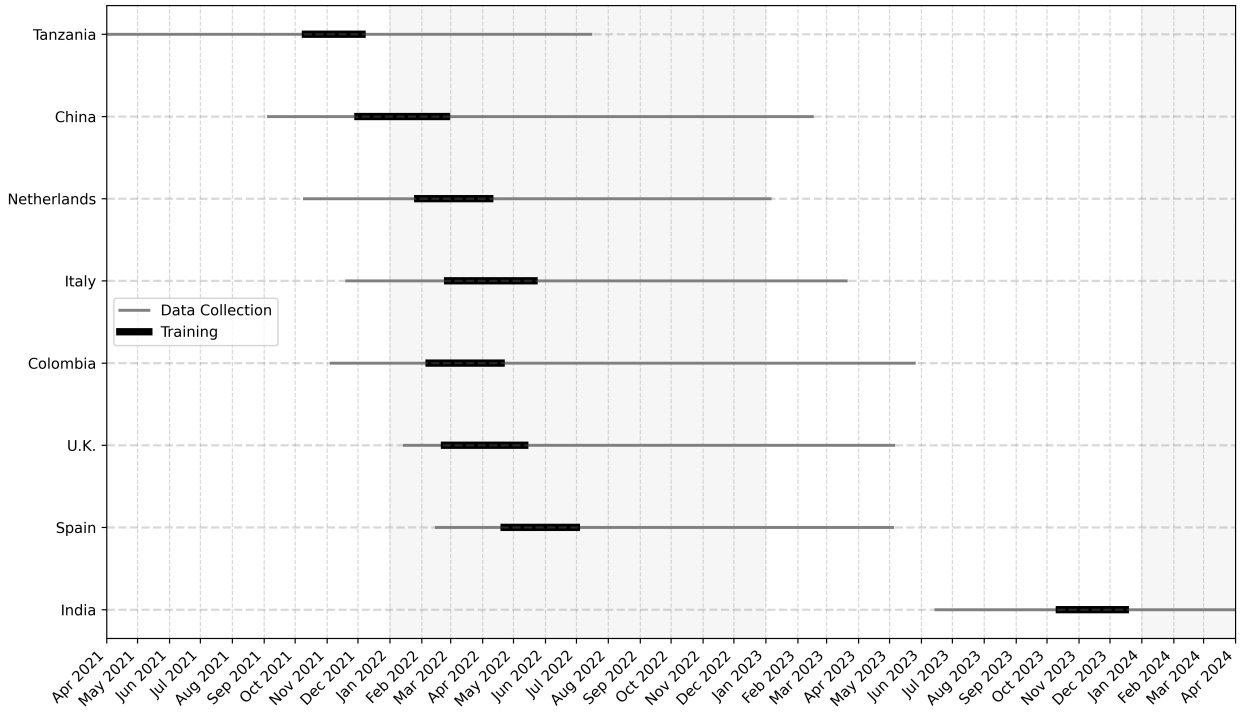


Figure I

This figure 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.

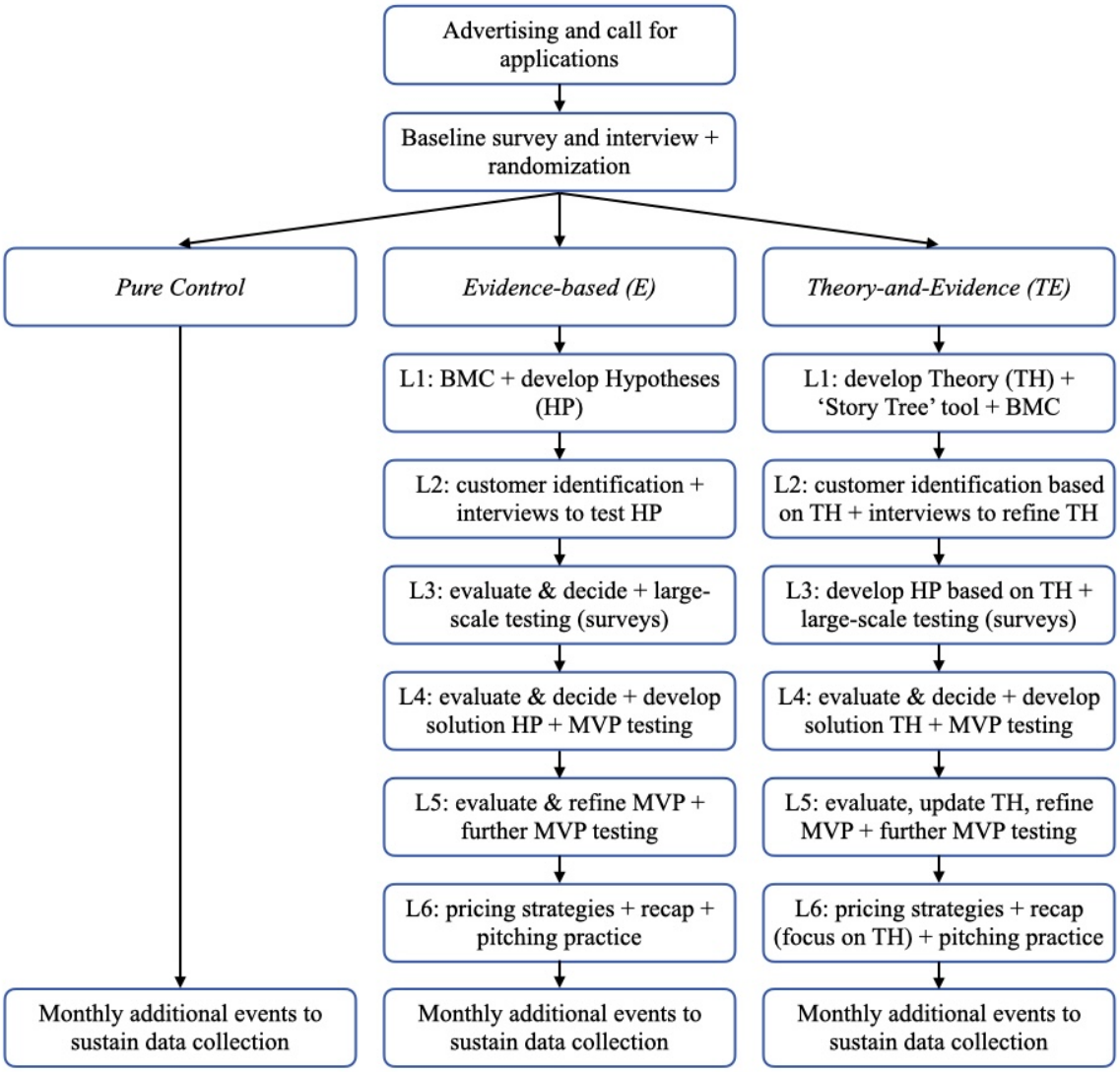


Figure II

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.

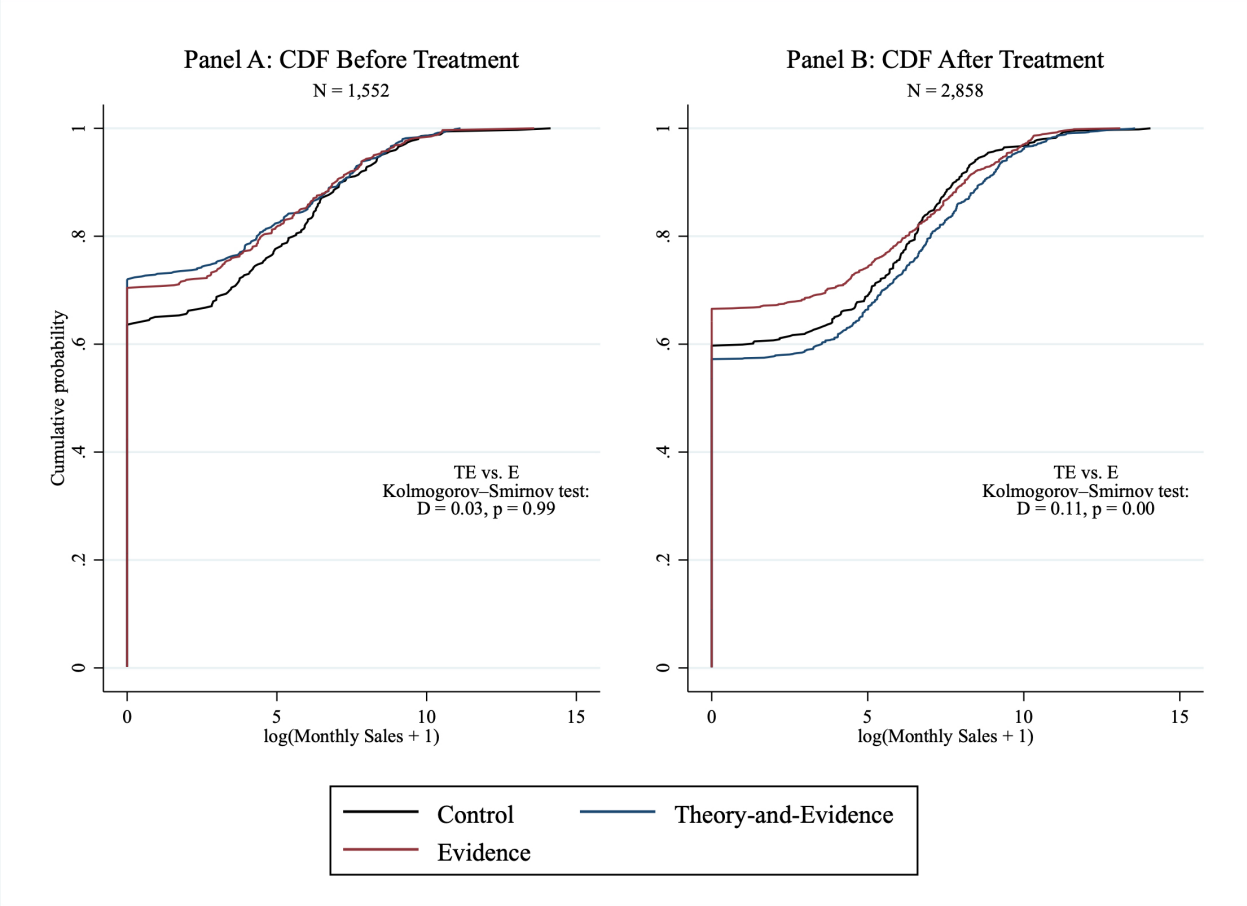


Figure III

This figure depicts the cumulative distribution functions (CDFs) of the $log(1 + Monthly\,Sales)$ outcome variable for startups across the three experimental arms. Panel A plots the CDFs for the baseline period, while Panel B plots the CDFs for the post-treatment periods. Both panels also report the results of two-sample Kolmogorov–Smirnov tests for equality of the distribution functions between the *TE* and *E* groups.

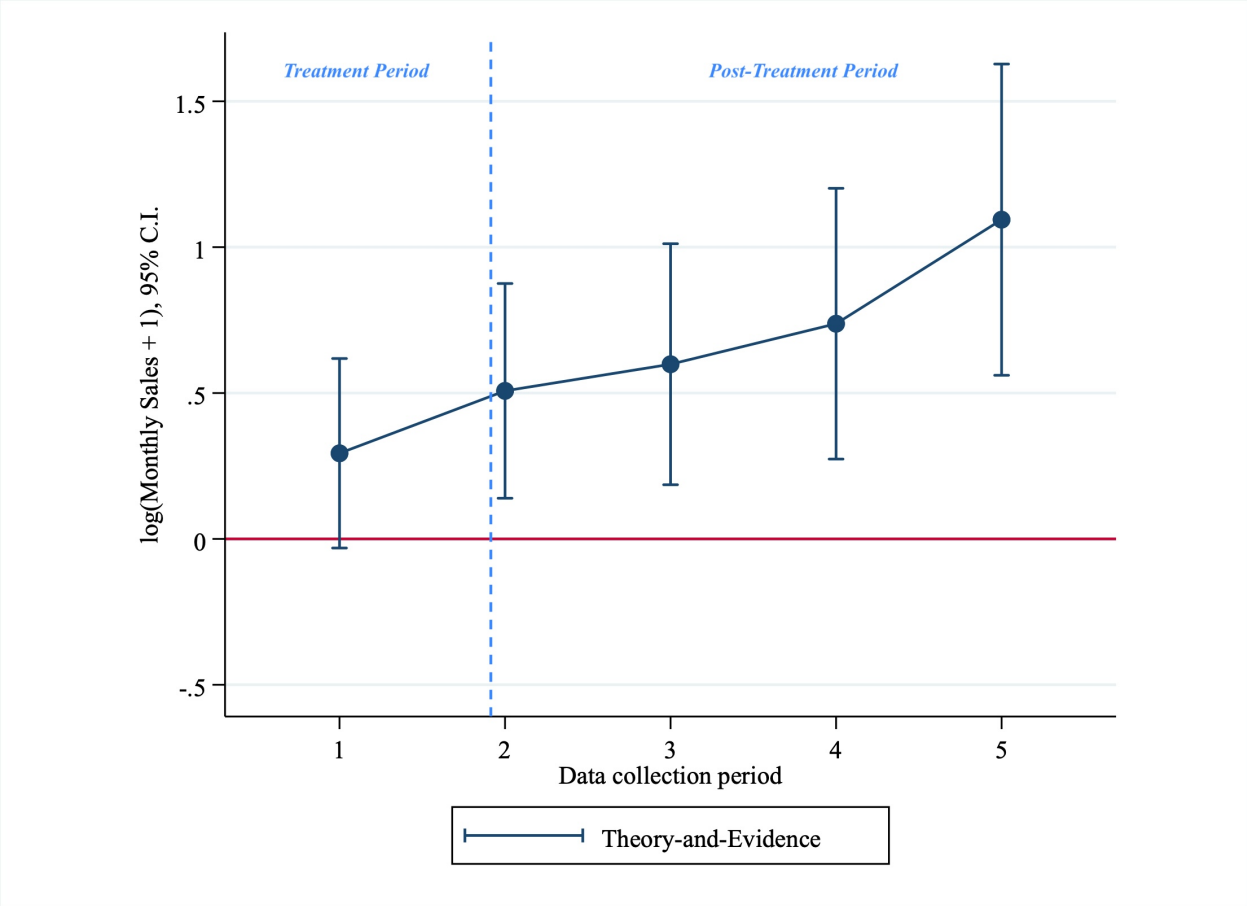


Figure IV

This figure graphs the coefficients of the ANCOVA regression on the logged measure of *Monthly Sales* on the subset of treated startups, using RCT $\times$ Period fixed effects, as well as standard errors clustered at the startup level. By interacting the treatment with the time period variable, we show the effect of the treatment 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.

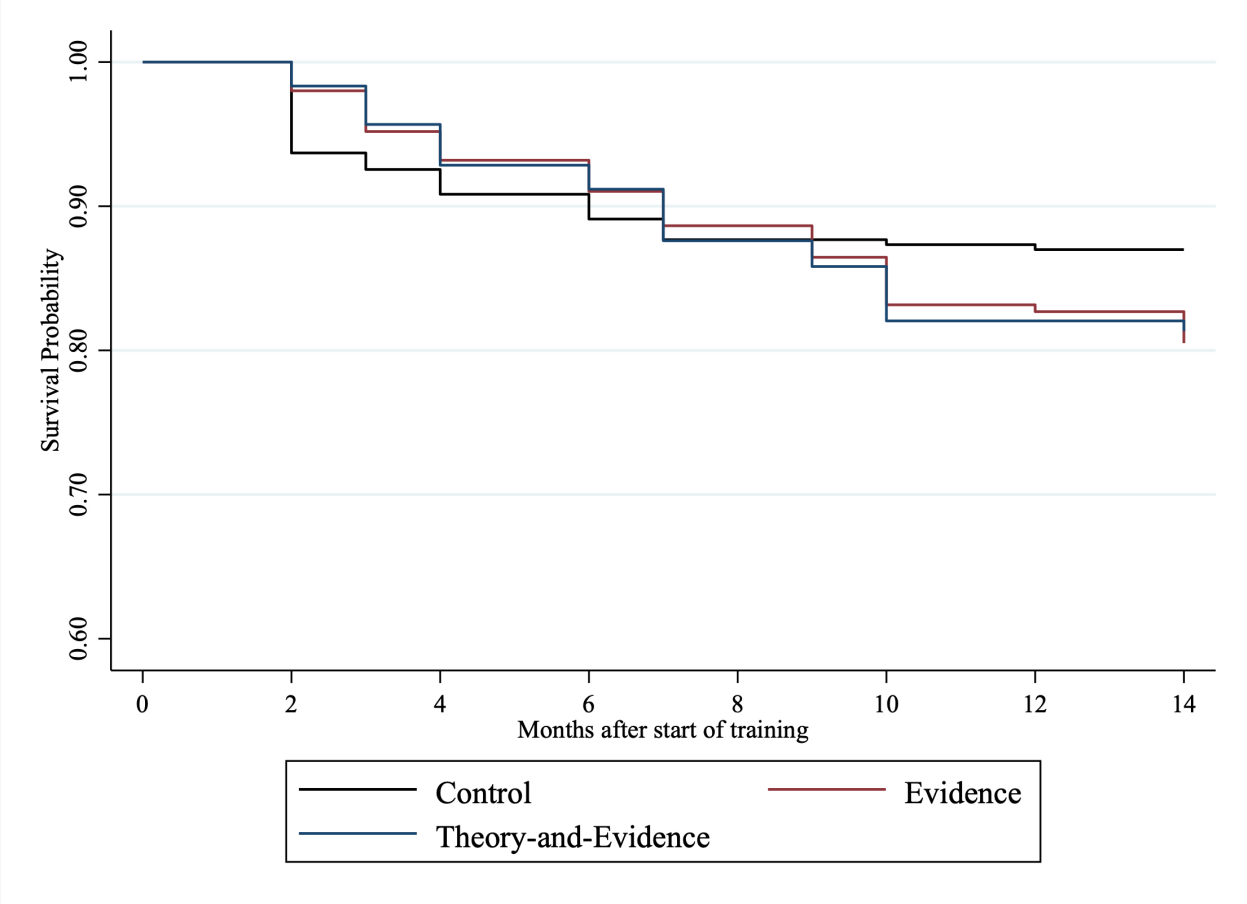


Figure V

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 ($\chi^2 = 0.00, p = 0.964$).

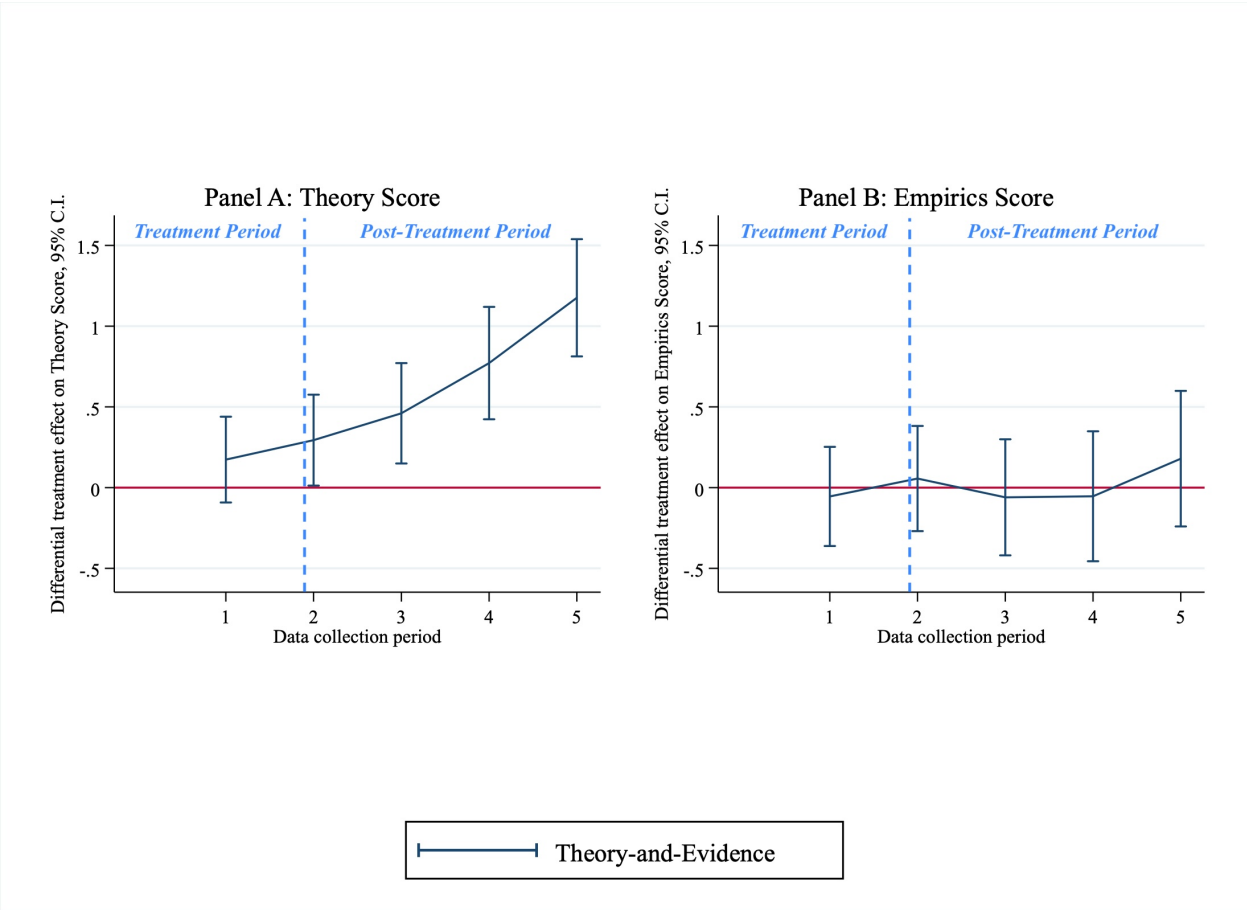


Figure VI

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 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	13.02
2	China	Shanghai	166	10.70
3	The Netherlands	Rotterdam	175	11.28
4	Italy	Milan	254	16.37
5	Colombia	Bogotá	206	13.27
6	U.K.	London	207	13.34
7	Spain	Barcelona	147	9.47
8	India	Chennai	195	12.56
Total			1,552	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
Monthly Sales[s,i]	Startup monthly sales reported in 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
Baseline Summary Statistics - Full Sample

	N	Mean	Median	Std. Dev.	Min	Max
Monthly Sales	1,552	2,805.11	0.00	42,744.91	0	1,382,333
Startup Employees	1,552	4.44	2.00	9.29	1	100
Months of Operations	1,552	15.55	7.00	21.52	0	268
Startup Sells a Product	1,552	0.26	0.00	0.44	0	1
Startup Sells a Service	1,552	0.36	0.00	0.48	0	1
Startup Sells Product and Service	1,552	0.39	0.00	0.49	0	1
Startup is For-Profit	1,552	0.92	1.00	0.27	0	1
Perceived Prob. of Termination	1,552	17.28	7.00	24.90	0	100
Age of founder	1,552	33.15	32.00	9.34	18	72
Gender (=1 Female)	1,552	0.36	0.00	0.48	0	1
Level of Education (1-3)	1,552	2.84	3.00	0.38	1	3
Econ or Business Degree	1,552	0.31	0.00	0.46	0	1
Working Full Time	1,552	0.40	0.00	0.49	0	1
Hours Worked per Week	1,552	24.48	19.00	22.12	0	100
Work Experience (Years)	1,552	9.41	7.00	8.38	0	40
Development Stage (1-5)	1,550	2.52	2.00	1.67	1	5
Theory Score	1,550	3.45	3.60	0.95	1	5
Empirics Score	1,550	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		
Monthly Sales	868.39 (220.11)	677.45 (141.38)	788.62 (161.37)	0.64	0.85	0.76
Startup Employees	3.69 (0.35)	4.80 (0.43)	4.52 (0.38)	0.06	0.22	0.36
Months of Operations	15.19 (1.05)	14.26 (0.82)	17.04 (0.97)	0.52	0.25	0.03
Startup Sells a Product	0.23 (0.02)	0.26 (0.02)	0.26 (0.02)	0.56	0.36	0.80
Startup Sells a Service	0.38 (0.03)	0.37 (0.02)	0.33 (0.02)	0.92	0.23	0.09
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	16.64 (1.25)	18.31 (1.07)	16.62 (1.00)	0.57	0.64	0.22
Age of founder	34.09 (0.48)	33.01 (0.38)	32.76 (0.39)	0.41	0.19	0.76
Gender (=1 Female)	0.39 (0.03)	0.38 (0.02)	0.33 (0.02)	0.69	0.30	0.07
Level of Education (1-3)	2.85 (0.02)	2.84 (0.02)	2.84 (0.02)	0.22	0.38	0.83
Econ or Business Degree	0.36 (0.03)	0.29 (0.02)	0.31 (0.02)	0.39	0.76	0.50
Working Full Time	0.34 (0.03)	0.41 (0.02)	0.43 (0.02)	0.64	0.34	0.47
Hours Worked per Week	22.31 (1.06)	24.36 (0.91)	25.87 (0.94)	0.37	0.04	0.20
Work Experience (Years)	10.20 (0.43)	9.31 (0.35)	9.06 (0.34)	0.51	0.22	0.72
Development Stage (1-5)	2.56 (0.09)	2.49 (0.07)	2.52 (0.07)	0.43	0.78	0.77
Theory Score	3.37 (0.05)	3.47 (0.04)	3.46 (0.04)	0.25	0.22	0.81
Empirics Score	2.23 (0.07)	2.41 (0.05)	2.33 (0.05)	0.24	0.82	0.29
F-test of joint significance (F-stat)				1.05	0.99	1.12
Observations	349	601	602	950	951	1,203

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 19, 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 stratified by location.

TABLE V

Treatments' effect on startup monthly sales

	Log	Raw	Winsorized, 95th	Winsorized, 90th	Sales Dummy
	(1)	(2)	(3)	(4)	(5)
Theory-and-Evidence	0.63***	0.37	0.47**	0.51***	0.07**
	(0.21)	(0.30)	(0.20)	(0.18)	(0.03)
Evidence	-0.06	-0.48*	0.11	0.20	0.01
	(0.21)	(0.25)	(0.24)	(0.20)	(0.03)
Log baseline sales	0.68***	0.68***	0.39***	0.30***	
	(0.07)	(0.06)	(0.04)	(0.04)	
Zero baseline sales	-0.36	2.55***	0.77*	0.26	
	(0.44)	(0.54)	(0.41)	(0.34)	
Constant	1.20***	4.17***	5.73***	5.97***	0.13***
	(0.43)	(0.62)	(0.45)	(0.37)	(0.02)
$\beta_{TE} = \beta_E$ (p-value)	0.000	0.002	0.055	0.041	0.012
Observations	2,858	2,858	2,858	2,858	1,078
Fixed Effects	RCT × Period	RCT × Period	RCT × Period	RCT × Period	RCT
Estimation	OLS	Poisson ML	Poisson ML	Poisson ML	OLS

Notes. This table reports ANCOVA estimates of the effect of the treatments on startup monthly sales. In Column (1), the outcome is $\log(1 + \text{Monthly Sales})$, which accommodates observations equal to zero. In Column (2), the outcome is the level of monthly sales (unwinsorized). In Columns (3) and (4), the outcome is the level of monthly sales winsorized (by RCT-period) at the 95th and 90th percentiles, respectively; the lower tail is not winsorized given the mass of observations at zero. In Column (5), the outcome is a binary measure indicating whether a startup generated positive sales during our study's time frame. Each column also reports the two-sided p-value of Wald tests for the equality of the two treatment coefficients. Standard errors (in parentheses) are clustered at the startup level in all columns. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE VI
Employment and Productivity Analyses

	Employees	Log(Sales / Employees)	TFP
	(1)	(2)	(3)
Theory-and-Evidence	0.10	0.56***	0.53***
	(0.39)	(0.20)	(0.14)
Evidence	0.53	-0.12	-0.14
	(0.41)	(0.20)	(0.14)
Baseline Employees	0.69***		
	(0.11)		
Baseline Log(Sales/Employees)		0.67***	
		(0.03)	
Baseline TFP			0.65***
			(0.02)
Constant	1.76***	0.57***	-0.52*
	(0.49)	(0.17)	(0.29)
$\beta_{TE} = \beta_E$ (p-value)	0.274	0.000	0.000
Observations	2,746	2,746	2,746
Fixed Effects	RCT × Period	RCT × Period	RCT × Period

Notes. This table reports ANCOVA estimates of the effect of the treatments on employment and productivity. In Column (1), the employment outcome is the total number of people working for the startup at each data collection point. In Column (2), the productivity outcome is the log of the ratio of startup monthly sales (in euros) to the total number of employees. In Column (3), the productivity outcome is a proxy for Total Factor Productivity (TFP), constructed as the residual from an OLS regression of Log(1 + Monthly Sales) on Log(Employees), controlling for RCT location × period fixed effects. Each column also reports the two-sided p-value from a Wald test of the equality of the two treatment coefficients. Standard errors (in parentheses) are clustered at the startup level in Columns (1) and (2); Column (3) reports bootstrap standard errors (200 replications), since the dependent variable is itself a generated regressor. The number of observations is lower than in Table V because the employment variable was collected via surveys only, whereas the sales variable was collected via both surveys and interviews at each data collection point; as a result, employment has more missing observations than sales. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE VII
Termination Analyses

	Termination Dummy		Termination Month
	(1)	(2)	(3)
Theory-and-Evidence	0.05**	0.05**	-2.84**
	(0.02)	(0.02)	(1.38)
Evidence	0.05**	0.05**	-2.83**
	(0.02)	(0.02)	(1.38)
Constant	0.19***		31.30***
	(0.04)		(1.87)
Observations	1,552	1,552	1,552
RCT Controls	Yes	Yes	Yes
Estimation	OLS	Logit	Tobit

Notes. This table reports cross-sectional analyses of termination (i.e., startup closure) for all startups in our sample. In the first two 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 Logit average marginal effects for robustness. Column (3) instead 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

TABLE VIII

Associations between performance and Theory and Experimentation scores

	Monthly Sales (Log)	Employees	Log(Sales/Empl.)	TFP
	(1)	(2)	(3)	(4)
Theory Score	0.08**	0.03	0.06**	0.06**
	(0.03)	(0.07)	(0.03)	(0.02)
Empirics Score	0.00	0.12*	-0.01	-0.02
	(0.03)	(0.07)	(0.03)	(0.02)
Baseline Theory Score	-0.08	0.17	-0.13	-0.14*
	(0.10)	(0.20)	(0.09)	(0.08)
Baseline Empirics Score	0.10	0.09	0.10	0.10*
	(0.08)	(0.18)	(0.07)	(0.05)
Log baseline sales	0.70***			
	(0.07)			
Zero baseline sales	-0.30			
	(0.45)			
Baseline Employees		0.68***		
		(0.12)		
Baseline Log(Sales/Empl.)			0.70***	
			(0.03)	
Baseline TFP				0.67***
				(0.02)
Observations	2,548	2,503	2,503	2,503
RCT × Period FE	Yes	Yes	Yes	Yes

Notes. This table reports ANCOVA regressions examining the correlation between theory and empirics scores and startup key outcomes. In Column (1), the outcome is $\log(1+\text{Monthly Sales})$; in column (2), the outcome is the total number of people working for the startup; in Column (3), the productivity outcome is the log of the ratio of startup monthly sales (in euros) to the total number of employees; in Column (4), the productivity outcome is the same TFP proxy employed in Table VI. Standard errors (in parentheses) are clustered at the startup level in columns (1) through (3); column (4) reports bootstrap standard errors (200 replications), since the dependent variable is itself a generated regressor. The number of observations in columns (2) through (4) is lower than column (1) because the employment variable was collected via surveys only, whereas the sales variable was collected via both surveys and interviews at each data collection point; as a result, employment has more missing observations than sales. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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, delivered 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 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).

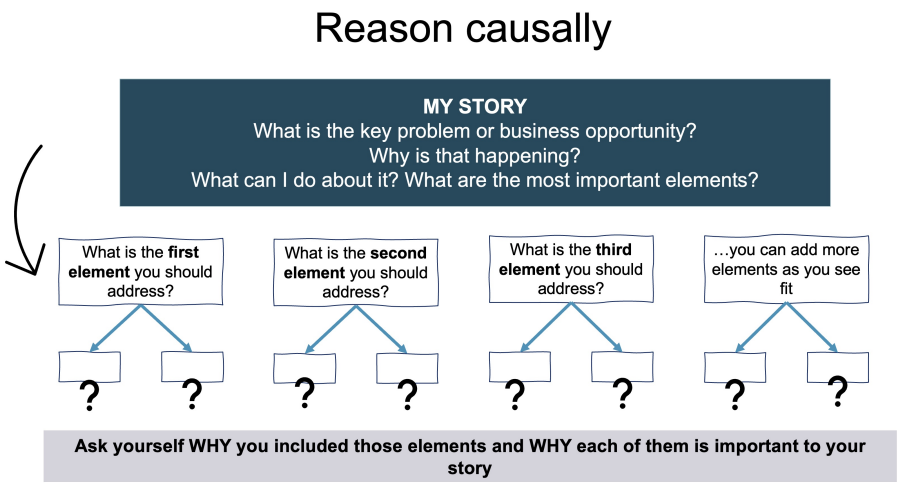


Figure A.1
The 'Story Tree' canvas nudging causal reasoning in theory-building, reproduced from Session 1 training slides of the *TE* treatment.

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, hierarchy 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 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 effect on the *Empirics Score* and on the *Theory Score*, respectively. We therefore estimate difference-in-differences regressions where the dependent variables are the *Theory* (Columns (1) and (2)) and *Empirics* scores (Columns (3) and (4)), and the DD coefficient indicates whether the startup i has taken the *TE* or *E* treatment at time t.

TABLE B.I
Treatment validity

	Theory Score		Empirics Score	
	(1)	(2)	(3)	(4)
Theory-and-Evidence × Post	2.13***	1.46***	2.76***	2.37***
	(0.34)	(0.32)	(0.37)	(0.36)
Evidence × Post	1.12***	0.77**	2.73***	2.27***
	(0.33)	(0.31)	(0.37)	(0.35)
Constant	8.97***	9.11***	5.70***	5.81***
	(0.09)	(0.09)	(0.10)	(0.10)
$\beta_{TE} = \beta_E$ (p-value)	0.000	0.006	0.939	0.758
Observations	5,198	5,198	5,198	5,198
RCT × Period 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. 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 RCT × Period fixed effects; 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 2.13 points ($\beta_{TE} = 2.13, p = 0.000$) compared to the control group, with the *E* treatment's increase being 1.12 points ($\beta_E = 1.12, p = 0.001$). Wald tests indicate that the two coefficients are statistically different from each other ($p = 0.000$). When tabulating the post-treatment means, the mean *Theory Score* of the control startups is 14.45, versus 16.00 for *E* startups and 16.98 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 ($\beta_{TE} = 2.76, p = 0.000$) and the *E* one ($\beta_E = 2.73, p = 0.000$) 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 ($p = 0.939$). Post-treatment means are equal to 8.41 for the control group, 11.33 for the *E* one, and 11.35 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 ($p = 0.931$).

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 and to Anthropic's latest 'Claude' API[14], 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 all the 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 0.00 and 0.01 for 'theory' and 'empirics' indices, 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 presents the empirical cumulative distribution functions of the RA and LLM scores for both the theory and empirics dimensions. Reassuringly, the three distributions track each other closely along the entire support of the scale. Focusing on OpenAI's model first, Figure B.1 indicates that GPT 5.4-mini assigns marginally higher scores than the RAs in the empirics dimension, but the practical agreement between human and model coding is high throughout. Wilcoxon signed-rank tests of the within-interview differences (RA minus GPT) reject the null of equality (theory: $z = -15.24$; empirics: $z = -26.92$; both $p < 0.001$; $N = 4,840$ pairs). These rejections, however, reflect the statistical power afforded by the large paired sample rather than any

[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 '5.4-mini' model from OpenAI and the 'Haiku 4.5' one from Anthropic, which were the latest available at the time of writing. We used enterprise accounts to preserve privacy and confidentiality of the data.

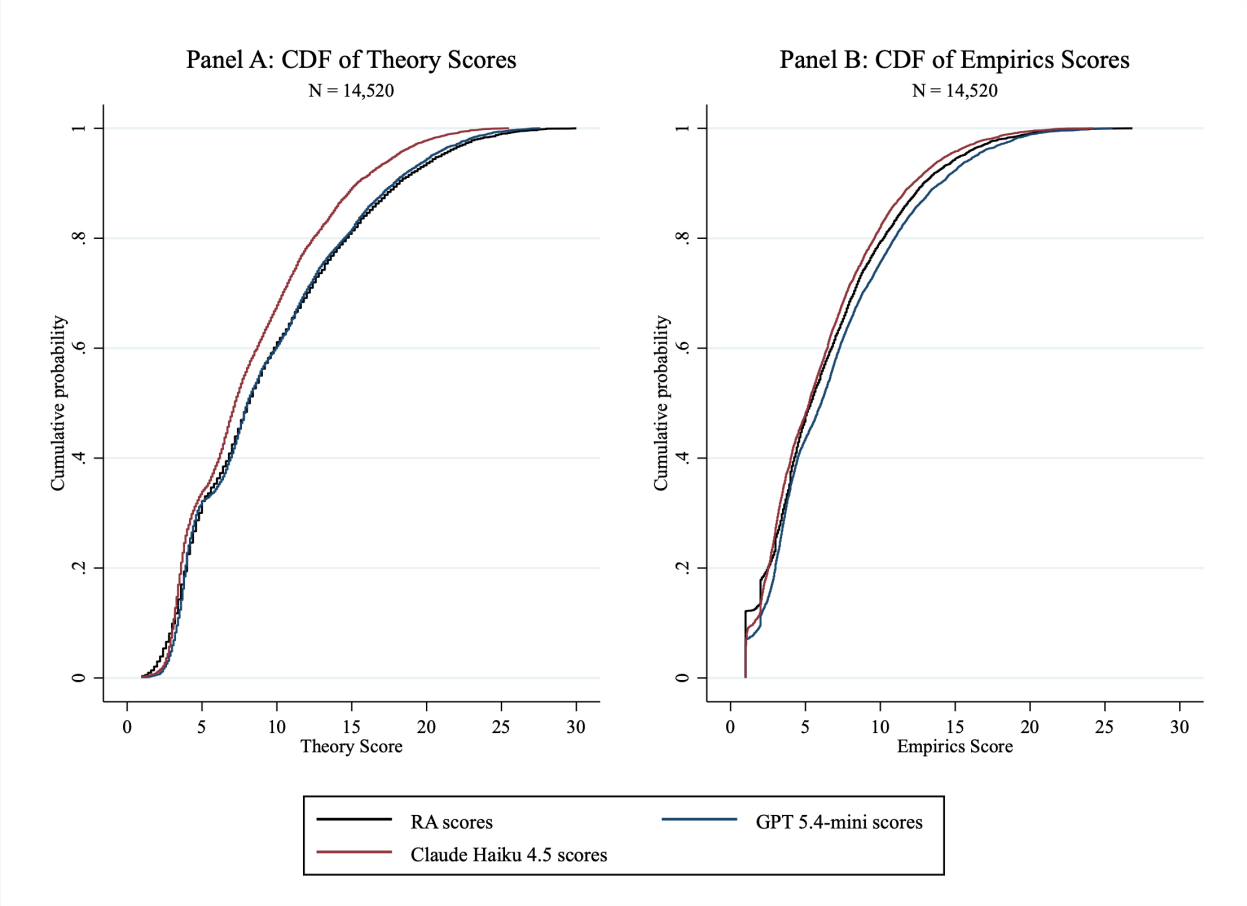


Figure B.1

This figure graphs the cumulative distribution functions (CDFs) of the *Theory Score* (Panel A) and *Empirics Score* (Panel B) variables, comparing the human RA scores to the LLM-generated ones (GPT 5.4-mini and Claude Haiku 4.5, the latest suitable models available at the time of writing).

economically meaningful divergence, with the differences being negligible in magnitude. For theory scores, the mean and median differences are 0.04 and -0.10 points, respectively; for empirics scores, they are -0.55 and -0.29 points.

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 Anthropic's 'Claude Haiku 4.5' API, which was the latest version available at the time of writing. Figure B.1 also reports the results of this second LLM (average of two repeats, as above). The comparison among CDFs in Panel B highlights a close alignment between this second AI model and the RAs when evaluating *Empirics Score*. However, the model tends to differ more significantly on the evaluation of the *Theory Score* (Panel A), with the Claude LLM being on average more conservative.

Overall, these tests broadly support our human RA scores, highlighting that our careful training and monitoring of the RAs resulted in effective data collection, with the human RA scores sitting between the two AI models' ones. 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 correlational 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 the two LLMs mentioned above). 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 precisely identified.

TABLE B.II

Associations between performance and Theory and Experimentation scores

	Monthly Sales (Log)	Employees	Log(Sales/Empl.)	TFP
	(1)	(2)	(3)	(4)
Theory Score (3 sources)	0.07*	-0.07	0.08**	0.08**
	(0.04)	(0.08)	(0.04)	(0.03)
Empirics Score (3 sources)	0.00	0.15*	-0.02	-0.03
	(0.04)	(0.08)	(0.03)	(0.03)
Baseline Theory Score	-0.07	0.39	-0.14	-0.15
	(0.14)	(0.30)	(0.13)	(0.11)
Baseline Empirics Score	0.15	0.09	0.14	0.13*
	(0.10)	(0.21)	(0.09)	(0.07)
Log baseline sales	0.70***			
	(0.07)			
Zero baseline sales	-0.29			
	(0.45)			
Baseline Employees		0.68***		
		(0.12)		
Baseline Log(Sales/Empl.)			0.69***	
			(0.03)	
Baseline TFP				0.67***
				(0.02)
Observations	2,548	2,503	2,503	2,503
RCT × Period FE	Yes	Yes	Yes	Yes

Notes. This table replicates ANCOVA regressions of Table VIII, using theory and empirics scores calculated as the averages between the scores from two LLMs (with two evaluations each) and the human RAs. In Column (1), the outcome is $\log(1+\text{Monthly Sales})$; in column (2), the outcome is the total number of people working for the startup; in Column (3), the productivity outcome is the log of the ratio of startup monthly sales (in euros) to the total number of employees; in Column (4), the productivity outcome is the same TFP proxy employed in Table VI. Standard errors (in parentheses) are clustered at the startup level in columns (1) through (3); column (4) reports bootstrap standard errors (200 replications), since the dependent variable is itself a generated regressor. The number of observations in columns (2) through (4) is lower than column (1) because the employment variable was collected via surveys only, whereas the sales variable was collected via both surveys and interviews at each data collection point; as a result, employment has more missing observations than sales. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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 at each data collection period. 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)	1,552	1,182	979	777	590	512
% on surviving sample	100%	76%	83%	79%	76%	87%
# of terminations	-	44	61	71	54	12
% attrition (excl. terminations)	-	21%	12%	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, 51% and 52% lower than for the control group ($exp(\beta_{TE}) = 0.49, exp(\beta_E) = 0.48, p = 0.000$). 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)
Monthly Sales (Log)	0.02
	(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	0.22*
	(0.12)
Startup is For-Profit	0.03
	(0.20)
Perceived Prob. of Termination	-0.00
	(0.00)
Gender (=1 Female)	0.11
	(0.11)
Bachelor's Degree or higher	0.29**
	(0.14)
Working Full Time	0.11
	(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	1,550

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 control 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) when comparing them. 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 34% higher than those without one $(exp(\beta) = 1.34, p = 0.042)$, 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, Fitzgerald, 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 statistically significant predictors of attrition, as identified in Table C.II [15]. We then estimate the ITT effects on performance employing weighted ANCOVA regressions, where the sample weights are computed as the inverse of the propensity scores. Column (1) of Table C.III presents results that closely match the ones reported in Table V.

[15] In addition to the treatment assignment and college degree, we also prudently included the *"Startup Sells a Service"* variable which was barely significant at the 10% confidence level in Table C.II

TABLE C.III
Robustness tests with attrition countermeasures

	Log(1+ Monthly Sales)			
	(1)	(2)	(3)	(4)
Theory-and-Evidence	0.63***	0.76***	0.33**	0.45**
	(0.21)	(0.13)	(0.14)	(0.17)
Evidence	-0.05	0.36***	-0.10	-0.08
	(0.21)	(0.13)	(0.14)	(0.15)
Log baseline sales	0.68***	0.32***	0.76***	0.62***
	(0.07)	(0.06)	(0.05)	(0.05)
Zero baseline sales	-0.36	-0.50	-0.19	-0.24
	(0.44)	(0.34)	(0.32)	(0.34)
Constant	1.21***	0.61*	0.86***	1.60***
	(0.43)	(0.34)	(0.31)	(0.48)
$\beta_{TE} = \beta_E$ (p-value)	0.000	0.001	0.001	0.001
Observations	2,858	5,818	5,818	5,818
Fixed Effects	RCT × Period	RCT × Period	RCT × Period	RCT × Period
Estimation Panel	Full, IPW	Zero-filled	LOCF-filled	MI-filled

Notes. This table reports ANCOVA robustness analyses of the effect of the treatments on startup monthly sales, accounting for the impact of attrition. In column (1), we weigh observations by the inverse of the propensity score, estimated with a nonparametric kernel regression. For column (2), missing values in the dependent variable are treated as having zero sales. For column (3), we replace missing values with the last available observation for each attritor (LOCF). In column (4), we employ predictive mean matching (PMM) to generate 20 imputations, drawing each missing value from its five nearest observed neighbors. Each column also reports the two-sided p-value of Wald tests for the equality of the two treatment coefficients. Standard errors (in parentheses) are clustered at the startup level in all columns. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

There is an ongoing debate in the literature regarding whether missing 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, as an additional check, we follow common practice in the literature (e.g., Anderson and McKenzie (2022)) and re-estimate our main specification on a panel in which the missing sales of attriting firms are set to zero. This imputation is motivated by the nature of attrition in our setting: because a substantial share of young firms leave the sample by ceasing operations,

a firm that no longer reports sales often could have a true value of zero rather than a genuinely missing one. Assigning the lowest admissible value of a nonnegative outcome to every attritor is deliberately conservative, and it corresponds to the lower-extreme imputation that underlies the worst-case bounds of Horowitz and Manski (2000). We therefore interpret the resulting estimates not as a point correction for attrition but as an additional robustness check. Reassuringly, the estimated effects in Column (2) of Table C.III remain positive and statistically significant under this imputation, although the extreme imputation is likely inflating the effect size vis-à-vis the control group that experiences greater attrition.

As an alternative approach, we then 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. This is a simple imputation technique that allows us to potentially incorporate information on the outside option value of entrepreneurs who terminate, as discussed in our model in Section 2.1. Even in this case, results are still positive and statistically significant, although likely attenuated by the strong assumptions behind the LOCF imputation. In particular, estimates in Column (3) of Table C.III indicate that startups in the *TE* treatment generated 39% higher monthly sales than the control group ($exp(\beta_{TE}) = 1.39, p = 0.020$).

Given that LOCF is a relatively naïve imputation approach, we perform a further robustness analysis using multiple imputation (MI, Little and Rubin, 2019; Imbens and Rubin, 2015), which are a potentially relevant solution to the problem of attrition in longitudinal studies with drop-outs (Little and Yau, 1996). Specifically, we employ predictive mean matching (PMM) to generate 20 imputations, drawing each missing value from its five nearest observed neighbors. Because PMM imputes only from observed values, it preserves the empirical distribution of our key *Monthly Sales* dependent variable and avoids implausible imputations. To predict the values to impute, we use treatment-by-period (RCT $\times$ Period) dummies – mirroring our main regression specification – and lagged sales values, which capture the serial persistence of firm-level monthly sales. MI estimates from column (4) of Table C.III confirm large and significant effects of the *TE* treatment, with startups in said group displaying 56% higher monthly sales compared to control ones ($exp(\beta_{TE}) =$

$1.56, p = 0.010$). Importantly, across all four specifications of Table C.III, Wald tests for the equality of *TE* and *E* coefficients[16] strongly reject the null hypothesis.

[16]Column (4) displays results from linear combination tests for the equality of the two treatment coefficients across the 20 imputations.

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 *Monthly Sales* from Columns (1), (3), and (5) 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 treatment effect for all three *TE* coefficients, but fail to reject the null for the *E* treatment, confirming our results from Table V.

TABLE D.I
FRI test results of Table V key coefficients

Statistic	T_{obs}	c	n	$p = c/n$	$SE(p)$	Interval
Log-OLS						
β_{TE}	0.627	1	2,000	0.001	0.001	[0.000, 0.003]
β_{E}	-0.055	1,506	2,000	0.753	0.010	[0.733, 0.772]
Poisson ML - winsorized 95th						
β_{TE}	0.467	80	2,000	0.040	0.004	[0.032, 0.049]
β_{E}	0.112	1,294	2,000	0.647	0.011	[0.626, 0.668]
Sales Dummy						
β_{TE}	0.071	29	2,000	0.015	0.003	[0.010, 0.021]
β_{E}	0.009	1,490	2,000	0.745	0.010	[0.725, 0.764]

Notes. T_{obs} represents the observed test statistic (i.e., β coefficients from Columns (1), (3), and (5) 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.

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.

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 (FWER). Across 2,000 resamples, 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, we focus on the comparison between *TE* and *E* treated groups – which is at the core of our paper. Column (1) of Table D.II reports the unadjusted p-values of the ANCOVA regressions described in equation

(13), comparing the *TE* and *E* treated groups. 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. The main adjusted p-values for our revenue and productivity outcomes, although often inflated by the MHT adjustments, remain statistically significant, suggesting that our results were not due to false rejections of the null hypotheses. The Poisson p-values are instead inflated by the adjustments above the conventional levels of significance, although we might be incurring in power limitations under this specification.

TABLE D.II
MHT Adjustments – Romano-Wolf stepdown

Outcome Variable	Unadjusted p-value	Romano-Wolf p-value	Holm p-value
	(1)	(2)	(3)
Log(Monthly Sales + 1)	0.000	0.001	0.002
Monthly Sales, 95th Winsorized	0.057	0.144	0.164
Employees	0.273	0.154	0.154
Log(Sales/Employees)	0.000	0.001	0.002

Notes. This table shows adjusted p-values computed in Stata using the 'rwolf2' package from Clarke, Romano, and Wolf (2020), with 2,000 resamples. Column (1) reports unadjusted p-values of our ANCOVA regressions. 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, extreme gradient boosting) with 250 splits of the data, as suggested by Chernozhukov et al. (2025). 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 four covariates:

- Number of months that the startup has been operational for (i.e., startup age in months)
- 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.

The GenericML routine (Chernozhukov et al., 2025) produces several results. First, it estimates a Best Linear Predictor (BLP) of the conditional average treatment effect. Given that we drop the control group and restrict attention to the two active arms, the estimand is the effect of the *TE* treatment relative to the *E* treatment as the counterfactual. Consistent with our ANCOVA design in equation (13), the BLP estimates the following model:

$$Y_{it} = \beta_1^{\text{BLP}} \widetilde{D}_i + \beta_2^{\text{BLP}} \widetilde{D}_i \widetilde{S}_i + \text{controls}_{it} + \varepsilon_{it} \qquad \text{for } t > 1, \tag{16}$$

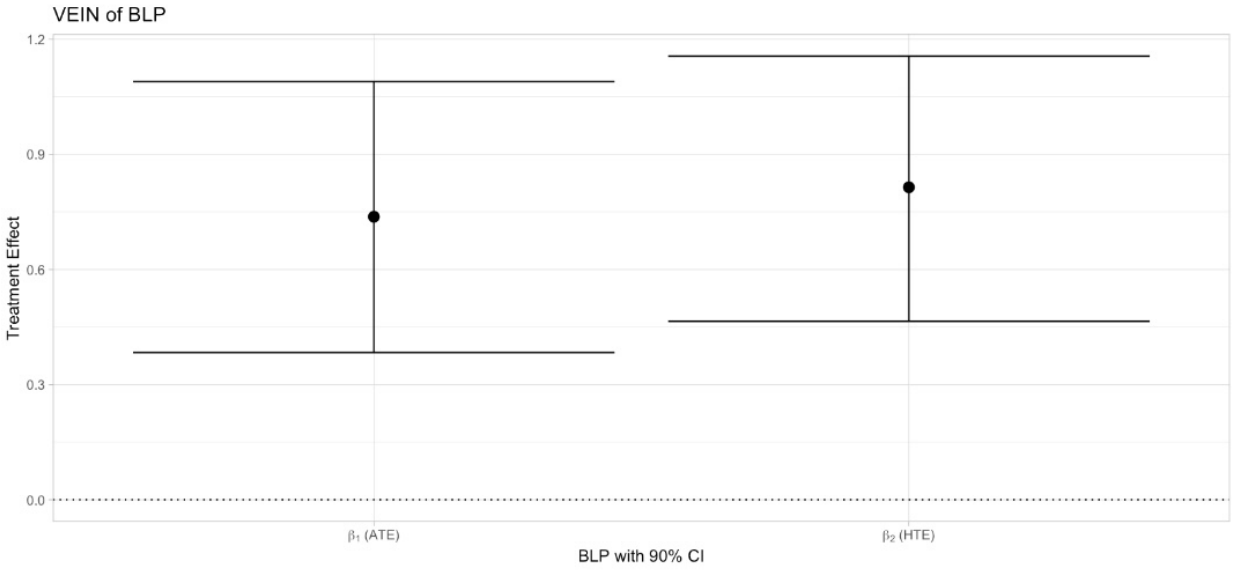
where Y_{it} is log monthly sales for startup i at period t; $\widetilde{D}_i = D_i - p_i$ is the treatment indicator (one for *TE*, zero for *E*) orthogonalized with respect to its known, RCT-stratified assignment probability p_i; and $\widetilde{S}_i = \hat{S}(Z_i) - \bar{S}$ is a demeaned machine-learning *proxy* for the conditional average treatment effect (CATE) as a function of the covariate vector Z_i. The term controls_{it} collects the RCT-by-period fixed effects λ_{rt} and a machine-learning proxy $\hat{B}(Z_i)$ for the baseline outcome, which plays the same precision-improving role as the ANCOVA covariate Y_{iB} in equation (13). The proxies $\hat{S}$ and $\hat{B}$ are learned on a random half of the sample while (16) is fit by weighted least squares on the other half; the split is repeated 250 times and we report the median estimates.[17]

The coefficient β_1^{BLP} estimates the average effect of the *TE* arm relative to the *E* arm. The coefficient β_2^{BLP} is our test for heterogeneity: if treatment effects are homogeneous, or the proxy fails to capture their variation, then $\beta_2^{\text{BLP}} \approx 0$ and the effect is summarized entirely by β_1^{BLP}; if the proxy captures genuine variation across startups, then $\beta_2^{\text{BLP}} > 0$. A significantly positive β_2^{BLP} is therefore direct evidence of heterogeneous *TE*-versus-*E* treatment effects.

Panel A in Figure E.1 shows that the ATE is positive and statistically significant ($\beta_1^{\text{BLP}} = 0.74, p = 0.000$), as well as closely aligned with previous results in section 5.1. Moreover, the estimate of β_2^{BLP} 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,

[17] The proxies are estimated by the best-performing learner among lasso, elastic net, random forest, support vector machine, and gradient boosting, and p-values are adjusted for the variation across splits. Weights are $[\,p_i(1-p_i)\,]^{-1}$.

Panel A: BLP of *TE* treatment effect



Panel B: GATES of *TE* treatment effect

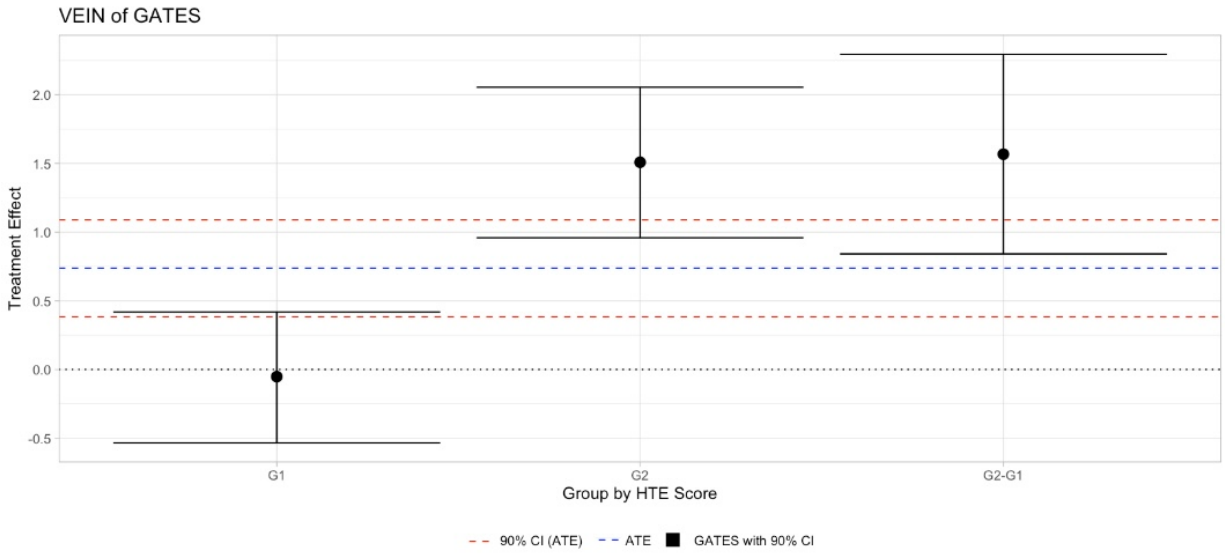


Figure E.1

This figure shows key plots compiled with the 'GenericML' R package by Chernozhukov et al. (2025). Panel A presents estimates of β_1^{BLP} (ATE) and β_2^{BLP} (HTE) from Equation (16). Panel B graphs the GATES using a median split across the distribution of treatment effects.

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 the GATE estimate for startups which have below-median treatment effects is practically null ($G_1 = -0.05, p = 0.822$). Conversely, the GATE estimate for startups with above-median effects is positive and highly statistically significant ($G_2 = 1.51, p = 0.000$). 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 then 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 seven months older than startups in the below-median group ($p = 0.000$). Given that most startups are extremely young, this result is natural and somewhat mechanical. CLAN results also signal a substantially higher proportion of females in the above-median group, with a difference of 9.3 percentage points ($\approx 30\%$) compared to the below-median group ($p = 0.001$). This interesting result is consistent with recent findings in the management literature, with Gagliardi and Novelli (2025) showing that "scientific" management treatments based on theory maps are particularly impactful for female founders by refining their value propositions. The two groups do not meaningfully differ in terms of years of work experience ($\Delta = 0.95, p = 0.056$), but there are on average 8% fewer founders with a college degree or higher among the above-median treatment effect group ($\Delta = -0.07, p = 0.001$). There might therefore be a potential partial substitution effect for those entrepreneurs who were already trained on causal reasoning in their university studies, for example in STEM subjects.

APPENDIX F: COST-BENEFIT ANALYSIS OF THE INTERVENTION

This appendix details the assumptions, data, and calculations underlying the cost-benefit assessment summarized in Section 6. We first describe the costing methodology (F.1) and the resulting program costs (F.2), then the monetization of benefits and a cost-benefit comparison with sensitivity analyses (F.3). Costs incurred in non-euro RCT locations were converted using the same ECB average exchange rates applied to the sales data (see Table II).

F.1 Costing Methodology

We follow the "ingredients method" of Dhaliwal et al. (2013) and the associated J-PAL costing guidelines. The objective of the exercise is to estimate the resources required to replicate the intervention that produced the effects estimated in Section 5. First, we exclude all research-related expenditures: research assistants' compensation, survey and interview administration, data-collection platforms, and the research team's time. We also exclude, in our baseline, the complementary events offered after the training period; these were designed to sustain participation in data collection, were offered to all three experimental arms – including the control group – and are therefore not part of the treatment contrast that identifies our estimates. Then, we include inputs that were donated or provided in kind at their market value: facilities made available by local partner institutions are priced at local commercial rental rates, and any unpaid time by the research team to train the instructors is priced at a reasonable market compensation rate. Because our performance estimates are intention-to-treat, both measures divide total costs by the number of startups assigned to the TE arm ($N_{TE} = 601$), so that costs and effects refer to the same population, irrespective of attendance.

F.2 Program Costs

Table F.I reports the full list of ingredients, their unit quantities and costs, and the implied cost per assigned startup. The program cost comprises the following components:

Curriculum development (one-time): the design of the training materials and take-home

exercises, amortized over the number of participants. The quantity of 140 hours for this task, used in Table F.I, was taken from the official agreement documents for the funding of the first RCT of our study; the amount of time to develop the ≈25% of *TE* contents was estimated more than proportionally at 45 hours, or roughly $1/3$ of the total, to account for their greater innovativeness.

Instructor training (per site): each of the 27 instructors received approximately 36 hours of standardized training (i.e., 1.5 hours for each hour of in-class teaching content) by a PhD researcher of the research team, valued at €25 per hour[18]. Instructor training was conducted online via Zoom, therefore no rooms were rented for the purpose. For each of the eight RCTs, all instructors were trained at the same time in group sessions (i.e., 36 hours for each of the eight RCTs), so that there would also be peer feedback and vicarious learning.

Delivery (instructor wages): each instructor taught six four-hour sessions per class of 20–25 startups, for a total of 24 in-class teaching hours (equal between *E* and *TE* treatments). To estimate the total instructor cost, we managed to recover the bank transfers of the payments received by 17 of the 27 total instructors[19]. We then divide this amount by 17 and multiply it by 27 to estimate the total instructor wages for the eight RCTs, which is then divided by two since instructors spent only half of their time teaching the *TE* curriculum.

Classrooms and teaching facilities: Classrooms were provided by host universities at no charge; we impute their opportunity cost at a rental-equivalent of €50 per hour, in line with published 2022 university facility rental rates for standard classrooms (e.g., FY2022–23 fee schedules at US public universities imply roughly €50 per hour for external users) and below commercial training-room rates. Each classroom was used for 24 hours (six 4-hour sessions), for an imputed cost of €1,200 per classroom. Applying a uniform rate across all eight sites overstates rental-equivalents in the lower-income sites (Colombia, India, Tanzania), making this a conservative

[18]Hourly compensation is not reported in the contracts; we impute a rate of €25 per hour (≈ $26.50 at the average exchange rate), consistent with the FY2022 NIH National Research Service Award stipend schedule for entry-level postdoctoral and predoctoral researchers (NIH Notice NOT-OD-22-132) and with the mean 2022 wage for social science research assistants in the BLS Occupational Employment and Wage Statistics (SOC 19-4061).

[19]For the Tanzania and China RCTs, we were not able to recover reliable payment data due to local specificities. Given that instructor wages in these countries were significantly lower than in the other high-income countries for which we have accurate data, the total amount in Table F.I likely overestimates the instructor costs, making our cost-benefit calculations more conservative.

(upper-bound) estimate. In total, 27 classrooms were employed throughout the eight locations, for an estimated total of €32,400.

Recruitment of participants (per site): the marketing campaigns, websites, and application screening required to enroll participants; following Dhaliwal et al. (2013) these targeting costs are program costs, not research costs, because any replication would incur them. To estimate these cost, we took the actual cost information for the Italy and U.K. RCTs, which had the most detailed financial reporting, and observed an average of €4,100 per location[20]. Our recruitment of participants was largely based on word-of-mouth advertising or social media posts by local partner institutions (and, in some Countries, by local government officials); however, monetary costs were incurred for hosting the program's websites, for running online advertising campaigns on social media, as well as for some specialized consulting to understand how to optimize our marketing efforts and contents for maximum effectiveness.

Miscellaneous (per site): this is a catch-all residual category, comprising various costs such as lunch boxes for the instructors and the assistants during training days, various consumables such as markers and water bottles, small gifts for the instructors at the end of their contracts, etc.

The resulting cost per assigned startup is

$$C_P = \frac{c^{tot}}{N_{TE}} = \frac{c^{curr} + c^{train} + c^{deliv} + c^{class} + c^{recr} + c^{misc}}{N_{TE}} = \frac{171,634}{601} = 285.58$$

Because curriculum development is a fixed cost and delivery costs are approximately linear in the number of participants, the per-startup cost at scale asymptotically declines to around €280 per participant. Finally, our experimental design permits an additional calculation that is arguably the most interesting: the *incremental* cost of the theory component. The *TE* and *E* arms shared identical contact hours, instructors, facilities, class sizes, and administration; they differed only in roughly 25% of the formal materials and in the framing of the shared content. Given that most business schools and online learning platforms already offer entrepreneurship courses that are

[20]Once again, applying this uniform rate across all eight sites overstates costs in the lower-income sites (Colombia, India, Tanzania), in line with our conservative approach at estimating costs.

roughly comparable to the *E* treatment (i.e., based on the Lean Startup approach, rapid testing and learning cycles, prototyping, etc.), the incremental ingredient is therefore the one-time adaptation of the curriculum and the corresponding portion of instructor training (roughly 50%, or 18 hours), returning a $c_\Delta \approx 8$ per startup at scale.

TABLE F.I
Program Cost Ingredients

Ingredient	Quantity	Unit cost	Total	Per assigned TE startup
Curriculum development	140 h	25 €/h	3,500	5.82
of which: TE adaptation	45 h	25 €/h	1,125	1.87
Instructor training	8 × 36 h	25 €/h	7,200	11.98
Instructor wages	27 × 24 h	N/A	83,734	139.32
Classrooms	27	1,200	32,400	53.91
Participant recruitment	8	4,100	32,800	54.58
Miscellaneous	8	1,500	12,000	19.97
Provider cost, C_P			171,634	285.58

Notes. Ingredients and unit costs follow the costing method of Dhaliwal et al. (2013). Research-related costs (research assistants, surveys, interviews, and post-training engagement events offered to all arms) are excluded; donated inputs are valued at market prices; costs from non-euro locations are converted at the ECB average exchange rates used for the sales data. Per-startup figures divide by the number of startups assigned to the *TE* arm ($N_{TE} = 601$), consistent with the intention-to-treat estimates in Table V. Site-level costs are pro-rated across the two treated arms.

F.3 Monetizing Benefits

Our experimental estimates identify the intention-to-treat (ITT) effect of the program on monthly startup sales of €970 per assigned *TE* startup. Sales, however, are not surplus: the benefit relevant for the cost–benefit analysis is the profit generated by treatment-induced sales, net of the variable inputs required to produce them. We therefore convert the sales effect into a profit effect by applying a marginal profit rate m, so that the monthly benefit per assigned startup is $m \times$ €970. Because the correct multiplier is the margin on incremental sales rather than the average accounting margin (which is contaminated by fixed costs), and because direct evidence on this parameter is scarce, we present results for a range of values of m.

We anchor this range in the experimental literature. In Anderson and McKenzie (2022),

control-group firms exhibit average profit-to-sales ratios of roughly 18–30 percent, while the ratio of their estimated treatment effect on profits to that on sales ($98 against $1,621 per month) implies a marginal rate of approximately 6 percent, albeit imprecisely estimated. We accordingly use $m \in \{5\%, 15\%, 30\%\}$, treating 15 percent as our central value, 5 percent as a conservative lower bound in line with the marginal ratio above, and 30 percent as an upper bound consistent with average margins among comparable small firms.

The second parameter is the persistence of the sales effect. Our data cover approximately twelve months following the end of treatment, over which the €970 monthly effect is estimated; benefits beyond that window are extrapolations that assume the effect persists at its estimated level. Table F.II reports the net present value (NPV) of benefits per assigned *TE* startup for horizons of 12 months (the observed window), 24 months, and 36 months, discounting monthly at an annual rate of 5 percent, a standard value in this literature (Alfonsi et al., 2020). Because both the €970 effect and the cost figures in Table F.I are expressed per *assigned* startup, the benefit and cost sides are consistent under the ITT interpretation, with non-compliance attenuating both symmetrically.

Even under the most conservative configuration – a 5 percent marginal profit rate and no persistence beyond the observed 12-month window – the discounted benefit of €567 per assigned startup exceeds the provider cost of €286 per assigned startup reported in Table F.I. Under the central scenario ($m = 15\%$, 12 months), benefits amount to €3,320 per assigned startup, more than eleven times the provider cost. We emphasize two caveats. First, these figures value private producer surplus and abstract from general-equilibrium effects: to the extent that treated startups' sales gains partially displace sales of untreated competitors, the social benefit is smaller than the private benefit (Anderson and McKenzie, 2022). Second, the extrapolated horizons assume no decay in the treatment effect; if effects fade, the 12-month row provides the relevant lower bound.

TABLE F.II
Benefit Sensitivity: NPV of Profit Gains per Assigned TE Startup

Horizon of sales effect	Marginal profit rate m		
	5% (€48.50/mo.)	15% (€145.50/mo.)	30% (€291.00/mo.)
12 months (observed)	567	1,701	3,401
24 months (extrapolated)	1,107	3,320	6,641
36 months (extrapolated)	1,621	4,863	9,726
Memo: provider cost per assigned startup, C_P		285.58	

Notes. Each cell reports the net present value, in base-year euros, of the monthly profit gain $m \times$ €970 per startup assigned to the *TE* arm, accumulated over the indicated horizon and discounted monthly at an annual rate of 5 percent. The 12-month horizon corresponds to the post-treatment window over which the sales effect is estimated; longer horizons assume the effect persists at its estimated level. The marginal profit rates span the range implied by control-group average margins (18–30 percent) and the marginal profit-to-sales ratio of treatment effects (≈6 percent) in Anderson and McKenzie (2022). The memo row reproduces the provider cost per assigned startup from Table F.I.