THE THEORY-BASED VIEW AND STRATEGIC PIVOTS: THE EFFECTS OF THEORIZATION AND EXPERIMENTATION ON THE TYPE AND NATURE OF PIVOTS

Jacob Valentine, Elena Novelli, and Rajshree Agarwal

APPENDIX

Please direct communications to Rajshree Agarwal at rajshree@umd.edu

TABLE OF CONTENTS

APPENDIX A. DETAILS ON RANDOMIZED CONTROL TRIAL

Table A1. Balance checks for the Treatment and Control Group Baseline Characteristics

To check that the randomization was successful we conducted a series of balance checks that compared treated and control entrepreneurs before the training on a variety of dimensions. They indicated the absence of significant differences between treated and control entrepreneurs at the baseline.

		Treatment		Control		Difference	
Variable	Elaboration	Mean	SD	Mean	SD	b	p
Business Age	Age of the business (years)	2.48	3.22	3.28	5.17	0.8	(0.14)
Team size	Number of team members	1.8	2.09	1.9	2.21	0.09	(0.72)
Gender (Female)	Proportion of women in the team	0.42	0.42	0.5	0.44	0.08	(0.12)
Age	Age (team average)	35.76	8.43	36.37	9.2	0.61	(0.58)
Hours - Total Weekly	Weekly hours dedicated to the company (team average)	31.51	18.29	29.61	17.12	-1.89	(0.39)
Background-Economics	Team members with Economics backgrounds (%)	0.14	0.29	0.15	0.29	0.01	(0.85)
Background - STEM	Team members with a STEM (Science Technology Engineering Mathematics) background (%)	0.29	0.39	0.36	0.43	0.07	(0.19)
Education	Highest educational level attained by team members (5 = PhD, 4 = MBA, 3 = MSc, 2 = BA, 1 = high school, 0 = other; team average)	2.67	0.8	2.58	0.79	-0.1	(0.33)
Confidence	Agreement on a 1–5 scale with the following statements (team average): "We are confident in our entrepreneurial skills", "We are sure we are deploying the best strategy for our business", "We are confident in our ability to manage our business", "We master the competences necessary for our venture", "We are sure there is no better business model for our idea"	3.41	0.69	3.34	0.76	-0.07	(0.44)
Probability Pivot Idea	Probability of making a radical change to the business	45.78	27.86	42.12	26.99	-3.66	(0.28)
Probability Pivot Problem	Probability of changing the problem and customer segment	38.13	25.86	40.55	26.26	2.43	(0.45)
Probability Expansion	Probability of expanding the business outside of the current industry or market	68.32	27.09	66.59	28.12	-1.73	(0.61)
Turnover Annual	Annual turnover (2018) £	57.14	166.3	83.13	226.26	25.99	(0.29)
Turnover Monthly	Monthly turnover (January 2019) £	5.81	20.26	7.04	28.29	1.24	(0.69)
Hours - % Innovation yearly	Working hours dedicated to the design of new products or services in the last year (2018, %)	45.92	32.98	40.02	32.68	-5.9	(0.15)
Hours - % Innovation monthly	Working hours dedicated to the design of new products or services in the last month (January 2019, %)	39.24	33.8	36.84	34.59	-2.41	(0.57)
Idea Value - Mean	Estimated value of the project (mean, 0 to 100)	66.83	16.8	66.62	20.22	-0.21	(0.93)
Idea Value - Range	Estimated value of the project (range, 0 to 100)	39.26	21.7	38	21.94	-1.26	(0.64)
Experience - Industry	Number of years of experience in industry (Team Average)	6.66	6.31	7.66	7.51	1	(0.25)
Experience - Work	Number of years of work experience (Team Average)	12.99	7.86	13.51	8.53	0.52	(0.61)
Experience - Entrepreneurial	Number of years of entrepreneurial experience (team average)	3.81	3.41	4.58	5.86	0.78	(0.20)
Experience - Managerial	Number of years of managerial experience (team average)	5.88	5.12	6.15	6.02	0.27	(0.69)
Observations		133		128		261	

Table A2: Scientific Intensity Components and Subcomponents

Research assistants were trained to code the level of scientific intensity of entrepreneurs building upon sixteen scores.

Component	Sub-component	Definition	Score
Theory	Clarity of theory	The extent to which the theory is understandable	0 (no theory) or from 1 (not clear) to 5 (extremely clear)
Theory	Articulation of theory	The extent to which the theory is detailed	0 (no theory) or from (not detailed) to 5 (extremely detailed)
Theory	Consideration of alternatives	The extent to which the theory includes alternative possible options	0 (no theory) or from 1 (no consideration of alternatives) to 5 (careful consideration of many alternatives)
Theory	Theory based on evidence	The extent to which the theory is based on objective evidence	0 (no theory) or from 1 (theory not based on objective evidence) to 5 (extremely based on objective evidence)
Hypotheses	Explicitness of hypotheses	The extent to which the respondent can articulate the fundamental assumptions that make his/her business viable	0 (no hypotheses) or from 1 (not explicit hypotheses) to 5 (extremely explicit)
Hypotheses	Coherence of hypotheses	The extent to which hypotheses are coherent with the theory	0 (no hypotheses) or from 1 (not coherent) to 5 (extremely coherent)
Hypotheses	Level of details of hypotheses	The extent to which hypotheses clearly indicate the details of what the entrepreneur wishes to learn and how to measure it	0 (no hypotheses) of from 1 (not detailed) to 5 (extremely detailed)
Hypotheses	Falsifiability of hypotheses	The extent to which it is possible to clearly determine (after tests) whether the hypotheses are supported or not	0 (no hypotheses) or from 1 (not falsifiable) to 5 (extremely falsifiable)
Tests	Coherence of tests	The extent to which the test is coherent with the hypotheses	0 (no tests) or from 1 (not coherent) to 5 (extremely coherent)
Tests	Validity of tests	The extent to which the test has been conducted in a context similar to which the business operates	0 (no hypotheses) or from 1 (not valid) to 5 (extremely valid)
Tests	Representativeness of tests	The extent to which the test has been conducted with a sample that is representative of the broad group the firm targets	0 (no hypotheses) or from 1 (not representative) to 5 (extremely representative)
Tests	Rigorousness of tests	The extent to which the appropriate test and procedure for that type of test have been chosen for hypotheses-testing	0 (no hypotheses) or from 1 (not rigorous) to 5 (extremely rigorous)
Evaluation	Data-based assessment	The extent to which the evaluation is based on data	0 (no hypotheses) or from 1 (not based on data) to 5 (extremely based on data)
Evaluation	Coherence of measures	The extent to which the measure used are consistent with the learning objective the entrepreneur has in mind	0 (no hypotheses) or from 1 (not coherent) to 5 (extremely coherent)
Evaluation	Systematic evaluation	The extent to which the evaluation is based on systematically collected and analyzed data	0 (no hypotheses) or from 1 (not systematic) to 5 (extremely systematic)
Evaluation	Explanatory power of evaluation	The extent to which the evaluation results in clarity on the main findings from the test and their implications for the business	0 (no hypotheses) or from 1 (not explanatory) to 5 (extremely explanatory)

Table A3 Attrition through the program

Some entrepreneurs withdrew from the program before the end of the observation period. Table 3 reports the number and percentage of entrepreneurs withdrawing from the program in each period.

Interview Number	In	Withdrawn	Withdrawn %
0	261	0	
1	223	38	15%
2	212	11	4%
3	207	5	2%
4	193	14	5%
5	185	8	3%
6	173	12	5%

7	163	10	4%
8	147	16	6%

Table A4. Attrition: Probability of withdrawing from the program

We checked that early withdraw from the program was not related to the allocation in the treatment condition with a regression estimating early withdraw as a function of the entrepreneur having been allocated in the treatment group. Results show that being allocated in the treatment group does not have a significant impact on the probability of early withdraw.

	(1)
	Early Withdraw
VARIABLES	OLS Cross-section
Treatment	-0.0167
	(0.7862)
Constant	0.4453
	(0.0000)
Observations	261
R-squared	0.0003
Clustered Errors	Firm

APPENDIX B: QUANTITATIVE ANALYSIS

Applying Natural Language Processing and ChatGPT to Develop Independent Measures of Scientific Intensity

Because all the interviews were recorded as audio files, it was necessary to first transcribe the audio files to prepare the textual information for coding by machine learning. We transcribed the audio files into text files by using OpenAI"s Whisper model, an automatic speech recognition system trained on 680,000 hours of various audio language data (Radford, Kim, Xu, Brockman, McLeavey, & Sutskever, 2023). Access to the model is provided without cost by OpenAI and can be used by anyone by installing via Github at https://github.com/openai/whisper.git. The Whisper model has four sizes, "tiny", "small", "medium", and "large", offering various speed and accuracy tradeoffs (Table B1). While the "tiny" model will run 32 times more quickly than the "large" model, its accuracy is dramatically lower. We utilized the large model for our transcription to maximize accuracy.

Table B1. Summary of the OpenAI Whisper model variants

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed
Tiny	39 M	tiny.en	tiny	~1 GB	~32x
Base	74 M	base.en	base	~1 GB	~16x
Small	244 M	small.en	small	~2 GB	~6x
Medium	769 M	medium.en	medium	~5 GB	~2x
Large	1550 M	N/A	large	~10 GB	1x

The transcription process was straightforward. After installing Whisper via Github and using NVIDIA"s CUDA toolkit for parallelization and the Google CoLab Pro+ Python environment, we looped through our directory of audio files and processed each audio file with the "Large" Whisper model. Any file that was not an audio file was immediately skipped. Any file that resulted in an error was also skipped. Less than 2% of the total number of files resulted in an error.[1] Some interviews were broken into multiple audio files for various reasons (e.g., the call was dropped). In these cases, we transcribed each file separately and combined the resulting text files afterward.

Once the interviews were transcribed, we cleaned and processed the transcripts with standard preprocessing techniques. We first removed common-language words (also called "stop words") such as "have", "was," "is," etc., single characters, special characters, and other linguistic artifacts that offer little to or detract from the NLP analysis. We did keep words that reflect causal logic, including "if," "then," and

[1] There are four fewer observations for machine-coded data due to missing/corrupted audio files

other relevant but commonly removed stop words. After removing stop words, we lemmatized and tokenized the transcripts using the Python library *nltk*. This step ensures that the same words are recognized regardless of their form; for example, "take" and "took" are recognized as the same word with the same meaning despite being spelled differently and having different tenses.

After the transcription process, we proceeded to develop independent dictionaries for measuring scientific intensity based on the language in the interviews by using ChatGPT via the web interface. Doing so improves the robustness of our analysis by providing an alternative measure of the scientificness of the entrepreneurs" language in the interviews from an unbiased source. We tested several prompts (the details of which are in Appendix A), and then used Term Frequency - Inverse Document Frequency (TF-IDF) analysis using the scikit learn library in Python (Elbagir & Yang, 2018) to determine the extent to which these dictionaries reflected the language used in the interviews. TF-IDF is a common statistical method used in natural language processing to measure how relevant a term is within a document relative to a collection of documents (i.e., relative to a corpus) by vectorizing the text (Ramos, 2003). After vectorizing the text with TF-IDF, we calculated the cosine similarity scores between the TF-IDF scores of both the transcripts and each dictionary to generate machine learning based scientific intensity scores.[2] We also used the BERT algorithm (Devlin, Chang, Lee, & Toutanova, 2018) to classify the text as a robustness check against the TF-IDF accuracy scores.

To create dictionaries for developing independent measures of scientific intensity, it was important to balance both independence and relevance. That is, we needed to (1) ensure that dictionaries were not created based on our prior knowledge as the researchers and also not based on any knowledge related to the treatment specifically, while also (2) ensuring that the dictionaries were sufficiently relevant to the corpus of interviews such that they could be useful in measuring scientific intensity. If we allowed ourselves to craft the dictionaries, they would naturally be susceptible to bias (even subconsciously). On the other hand, if the dictionaries were not relevant to the interviews, we would be unable to use them in analysis because they would not be able to identify meaningful information in corpus. Thus, to accomplish our first objective, we utilized ChatGPT with a variety of straightforward prompts to develop multiple dictionaries (one dictionary per SI dimension per prompt), and then, to accomplish our second objective, we measured

[2] We used Python to implement the TF-IDF analysis and run the cosine similarity scores. Our code is available upon request.

the relevance of each dictionary relative to the corpus with TF-IDF vectorization and cosine similarity scores (as discussed in the body of the paper). This appendix explains how we created and selected the dictionaries we ultimately used for our analyses.

The first prompt we used reflected the way we prompted the research assistants to code the interviews (Table 3). This was the most lengthy and restrictive of the prompts we used. This prompt and the resulting dictionaries are captured in Table B2.

Table B2. Dictionary 1: Detailed Based on Table 3	
Prompt	I am performing NLP text analysis on a set of documents to examine the extent to which they reflect scientific reasoning. Please generate four dictionaries or corpuses of words that I can use to assess the extent to which the documents reflect four different components of scientific reasoning: (1) scientific theorizing, including the extent to which the theory is understandable, the extent to which the theory is detailed, the extent to which the theory includes alternative possible options, and the extent to which the theory is based on objective evidence; (2) hypothesis development, including the extent to which the respondent can articulate the fundamental assumptions that make his/her business viable, the extent to which hypotheses are coherent with the theory, the extent to which hypotheses clearly indicate the details of what the entrepreneur wishes to learn and how to measure it, the extent to which it is possible to clearly determine (after tests) whether the hypotheses are supported or not; (3) experimentation or testing of the hypothesis, including the extent to which the test is coherent with the hypotheses, the extent to which the test has been conducted in a context similar to which the business operates, the extent to which the test has been conducted with a sample that is representative of the broad group the firm targets, the extent to which the appropriate test and procedure for that type of test have been chosen for hypotheses-testing; and (4) the evaluation of the results of the experiment, including the extent to which the evaluation is based on data, the extent to which the measure used are consistent with the learning objective the entrepreneur has in mind, the extent to which the evaluation is based on systematically collected and analyzed data, the extent to which the evaluation results in clarity on the main findings from the test and their implications for the business.

Theory Dictionary	scientific_theorizing_dict = {
	'understandability': ['clear', 'coherent', 'intelligible', 'comprehensible', 'accessible', 'lucid', 'transparent', 'explicit', 'digestible', 'straightforward'],
	'detail': ['elaborate', 'comprehensive', 'thorough', 'exhaustive', 'detailed', 'intricate', 'in-depth', 'profound', 'meticulous'],
	'alternative_options': ['diversified', 'varied', 'multiple_perspectives', 'alternatives_considered', 'diverse_possibilities', 'range_of_options', 'various_alternatives', 'pluralistic', 'different_angles', 'myriad_choices'],
	'objective_evidence': ['empirical', 'factual', 'observable', 'data-driven', 'evidence-based', 'fact-based', 'verifiable', 'tangible', 'substantiated', 'concrete']
	}

Hypothesis Dictionary	hypothesis_development_dict = { 'fundamental_assumptions': ['foundational_beliefs', 'core_assumptions', 'underlying_principles', 'premises', 'basic_tenets', 'central_hypotheses', 'fundamental_postulates', 'key_presumptions', 'primary_conjectures', 'principal_suppositions'], 'coherence_with_theory': ['consistent', 'aligned', 'harmonized', 'corresponding', 'congruent', 'synchronized', 'conforming', 'compatible', 'accordant', 'integrated'],

	'clarity_in_hypotheses': ['explicit', 'clearly_defined', 'precise', 'articulated', 'unambiguous', 'distinct', 'clearly_expressed', 'clearly_outlined', 'clearly_stated', 'clearly_formulated'], 'testability': ['measurable', 'observable_outcomes', 'quantifiable', 'testable_predictions', 'verifiable', 'empirically_testable', 'assessable', 'measurable_parameters', 'appraisable', 'observable_metrics'] }

Testing Dictionary	experimentation_dict = {
	'coherence_with_hypotheses': ['consistent', 'congruent', 'corresponding', 'aligned', 'compatible', 'synchronized', 'conforming', 'harmonized', 'accordant', 'integrated'],
	'context_similarity': ['contextual', 'relevant_environment', 'applicable_context', 'situation-specific', 'pertinent_setting', 'appropriate_circumstances', 'fitting_conditions', 'relevant_framework', 'scenario-based', 'environmental_congruity'],
	'representative_sample': ['representative', 'typical', 'inclusive', 'broad-based', 'generalizable', 'diverse', 'varied', 'comprehensive', 'all-encompassing', 'wide-ranging'],
	'appropriate_test_and_procedure': ['methodologically_sound', 'rigorous', 'valid', 'reliable', 'apt_methodology', 'rigorous_procedure', 'sound_method', 'validated_approach', 'credible_process', 'thorough_testing']
	}

Evaluation Dictionary	evaluation_results_dict = {
	'data_based_evaluation': ['data-driven', 'analytical', 'statistically_analyzed', 'empirically_assessed', 'quantitatively_evaluated', 'numerically_assessed', 'analytic_scrutiny', 'scientific_evaluation', 'statistical_scrutiny', 'numerical_analysis'],
	'consistency_with_learning_objectives': ['goal-oriented', 'objective-aligned', 'purposeful', 'targeted', 'directed_towards_objectives', 'intentional', 'goal-driven', 'objective-focused', 'purpose-driven', 'result-oriented'],
	'systematically_collected_and_analyzed_data': ['rigorous_analysis', 'systematic_collection', 'methodical_examination', 'analytical_approach', 'comprehensive_scrutiny', 'thorough_analysis', 'organized_data_collection', 'methodical_assessment', 'in-depth_scrutiny', 'rigorous_examination'],
	'clarity_on_main_findings': ['clearly_articulated_results', 'transparent_conclusions', 'conclusive', 'unambiguous_outcomes', 'clearly_stated_implications', 'explicit_findings', 'distinct_results', 'clearly_delineated_conclusions', 'definitive_outcomes', 'clearly_expressed_implications']
	}

The second prompt we used was the simplest of our prompts. We chose this prompt as an attempt to leave as much structure, information, and other potentially biasing or confounding factors out of the prompt as possible, leaving the structures and styles of the dictionaries to ChatGPT. This prompt and its resulting dictionaries are found in Table B3.

Table B3. Dictionary 2: Basic	
Prompt	I am performing NLP text analysis on a set of documents to examine the extent to which they reflect scientific reasoning. Please generate four dictionaries or corpuses of words that I can use to assess the extent to which the documents reflect four different components of scientific reasoning: (1) scientific theorizing, (2) hypothesis development, (3) experimentation or testing of the hypothesis, and (4) the evaluation of the results of the experiment. Give me at least 40 words per category.
Theory Dictionary	scientific_theorizing = [

	"theory", "hypothesis", "model", "framework", "paradigm", "concept", "postulate", "proposition", "explanation", "conceptualization", "construct", "speculation", "assumption", "conjecture", "principle", "law", "axiom", "theorem", "thesis", "idea", "generalization", "speculation", "reasoning", "deduction", "induction", "abstraction", "conceptual", "formulation", "speculation", "premise", "scientific rationale", "scientific model", "scientific framework", "scientific concept", "scientific principle", "scientific law", "scientific axiom", "scientific theorem", "scientific thesis", "scientific idea"]
Hypothesis Dictionary	hypothesis_development = ["hypothesis", "conjecture", "proposal", "supposition", "assumption", "prediction", "speculation", "educated guess", "inference", "proposition", "theory", "concept", "model", "idea", "tentative explanation", "working hypothesis", "testable statement", "empirical hypothesis", "research hypothesis", "null hypothesis", "alternative hypothesis", "independent variable", "dependent variable", "variable", "experimental hypothesis", "cause-and-effect relationship", "research question", "investigational hypothesis", "conjectural statement", "speculative hypothesis", "conceptual hypothesis", "formulated hypothesis", "proposed explanation", "anticipatory statement", "preliminary assumption", "assumed relationship", "speculative proposition", "theoretical assumption", "conditional proposition"]
Testing Dictionary	experimentation_testing = ["experiment", "test", "trial", "investigation", "study", "research", "analysis", "observation", "measurement", "data collection", "controlled study", "controlled experiment", "clinical trial", "field study", "laboratory investigation", "empirical research", "systematic observation", "manipulation of variables", "controlled observation", "field experiment", "laboratory experiment", "randomized controlled trial", "intervention study", "longitudinal study", "cross-sectional study", "empirical analysis", "scientific inquiry", "research design", "research methodology", "data gathering", "data analysis", "statistical analysis", "controlled conditions", "treatment group", "control group", "placebo group", "experimental group", "measurement instrument"]
Evaluation Dictionary	evaluation_results = ["analysis", "interpretation", "conclusion", "findings", "results", "outcome", "data", "evidence", "observation", "measurement", "assessment", "evaluation", "scrutiny", "examination", "validation", "verification", "scrutiny", "scrutiny", "interpretation", "statistical analysis", "statistical significance", "significance level", "confidence interval", "p-value", "correlation", "causation", "effect size", "reliability", "validity", "precision", "accuracy", "error analysis", "trend analysis", "pattern analysis", "graphical representation", "qualitative analysis", "quantitative analysis", "critical evaluation", "logical analysis", "systematic review", "meta-analysis"]

Our third prompt built on the second by offering only slight guidance in the definition of theorizing, which we believe is especially distinct in the realm of entrepreneurship and strategy (compared to traditional science). Whereas traditional science would suggest that theorizing consists largely of understanding the larger, existing body of knowledge and developing predictions based on the logical conclusions of that body of knowledge, a theory of value is focused on the logical conclusions of an entrepreneur's beliefs about the world and his or her idea in that world. This prompt and its resulting dictionaries are shown in Table B4.

Table B4: Dictionary 3: Guided for Theory	
Prompt	I am performing NLP text analysis on a set of documents to examine the extent to which they reflect scientific reasoning. Please generate four dictionaries or corpuses of words that I can use to assess the extent to which the documents reflect four different components of scientific reasoning: (1) scientific theorizing, especially the use of causal logic, problem-solution formulation, and theory development; (2) hypothesis development, (3) experimentation or testing of the hypothesis, and (4) the evaluation of the results of the experiment.

Theory Dictionary	scientific_theorizing_dict = {
	'Causal Logic': ['cause', 'effect', 'consequence', 'result', 'impact', 'influence', 'correlation', 'relationship', 'connection'],
	'Problem-Solution Formulation': ['problem', 'issue', 'challenge', 'solution', 'resolve', 'address', 'overcome', 'tackle', 'strategy'],
	'Theory Development': ['theory', 'model', 'framework', 'concept', 'construct', 'hypothesis', 'explanation', 'proposition', 'formulation']
	}

Hypothesis Dictionary	hypothesis_development_dict = {
	'Hypothesis': ['hypothesis', 'conjecture', 'proposition', 'assumption', 'thesis', 'supposition', 'speculation', 'prediction', 'postulate'],
	'Testing Language': ['test', 'experiment', 'trial', 'investigate', 'examine', 'assess', 'analyze', 'verify', 'validate', 'assess'],
	'Prediction': ['predict', 'anticipate', 'forecast', 'foresee', 'expect', 'projection', 'estimate', 'guess', 'envisage']
	}

Testing Dictionary	experimentation_dict = {
	'Experimental Methods': ['experiment', 'methodology', 'procedure', 'approach', 'design', 'setup', 'protocol', 'technique', 'method'],
	'Data Collection': ['data', 'measurement', 'observation', 'recording', 'collection', 'gather', 'obtain', 'acquire', 'record'],
	'Variables': ['independent variable', 'dependent variable', 'control variable', 'manipulate', 'vary', 'change', 'alter', 'modify']
	}

Evaluation Dictionary	evaluation_dict = {
	'Analysis': ['analyze', 'interpret', 'examine', 'assess', 'evaluate', 'scrutinize', 'inspect', 'review', 'critique'],
	'Conclusions': ['conclude', 'deduce', 'infer', 'derive', 'result', 'outcome', 'implication', 'significance', 'consequence'],
	'Statistical Terms': ['statistical significance', 'p-value', 'correlation coefficient', 'standard deviation', 'mean', 'median', 'variance', 'confidence interval']
	}

After the completion of these dictionaries, we combined all dictionaries within each component of SI to create a single aggregated dictionary for each of the four components. This did not require an additional prompt for ChatGPT. Instead, we took all of the previously created dictionaries, combined them, and removed any duplicate entries (all within the individual components).

For Dictionary 5, we utilized a prompt that would generate dictionaries specific to scientific intensity in the entrepreneurial context, supposing that it might better reflect the language used by our entrepreneurs. This prompt and its resulting dictionaries are listed in Table B5.

Table B5. Dictionary 5: Entrepreneurship-specific	
Prompt	I am performing NLP text analysis on a set of documents to examine the extent to which they reflect scientific reasoning. Please generate four dictionaries or corpuses of words that I can use to assess the extent to which the documents reflect four different components of scientific reasoning: (1) scientific theorizing, (2) hypothesis development, (3) experimentation or testing of the hypothesis, and (4) the evaluation of the results of the experiment. Give me approximately 30 words per category. Note that the context for these documents is interviews with entrepreneurs, so be sure that the words in each corpus reflect scientific reasoning in entrepreneurship (as opposed to contexts like research scientists) and are words that entrepreneurs might use as they approach their ventures scientifically.

Theory Dictionary	scientific_theorizing_corpus = {
	'hypothesis', 'framework', 'model', 'paradigm', 'conceptualize', 'formulate',
	'structure', 'construct', 'propose', 'conjecture', 'conceptual', 'theoretical',
	'framework', 'systematize', 'structure', 'conceptualization', 'paradigmatic',
	'conceptual framework', 'model', 'postulate', 'theoretical foundation',
	'conceptual model', 'hypothesis-driven', 'conceptual framework',
	'propose', 'conjecture', 'conceptual', 'theoretical', 'systematize',
	'structure', 'formulate', 'propose', 'conjecture', 'conceptual',
	'theoretical', 'systematize', 'conceptualization', 'paradigmatic',
	'conceptual framework', 'model', 'postulate', 'theoretical foundation',
	'conceptual model', 'hypothesis-driven', 'conceptual framework',
	'propose', 'conjecture', 'conceptual', 'theoretical', 'systematize'
	}

Hypothesis Dictionary	hypothesis_development_corpus = {
	'conjecture', 'assumption', 'propose', 'propose a hypothesis', 'hypothesize',
	'speculate', 'suggest', 'postulate', 'formulate', 'propose an idea',
	'form an idea', 'develop a theory', 'propose a conjecture', 'advance a hypothesis',
	'suggest a theory', 'put forward a hypothesis', 'conceptualize',
	'construct a hypothesis', 'generate a hypothesis', 'formulate a hypothesis',
	'propose a theory', 'make a conjecture', 'suggest an idea', 'theorize',
	'propose an explanation', 'suggest a hypothesis', 'develop an assumption',
	'make a supposition', 'hypothesize a theory', 'conjecture', 'assumption',
	'propose', 'propose a hypothesis', 'hypothesize', 'speculate', 'suggest',
	'postulate', 'formulate', 'propose an idea', 'form an idea', 'develop a theory',
	'propose a conjecture', 'advance a hypothesis', 'suggest a theory',
	'put forward a hypothesis', 'conceptualize', 'construct a hypothesis',
	'generate a hypothesis', 'formulate a hypothesis', 'propose a theory',
	'make a conjecture', 'suggest an idea', 'theorize', 'propose an explanation',
	'suggest a hypothesis', 'develop an assumption', 'make a supposition',
	'hypothesize a theory'
	}

Testing Dictionary	experimentation_corpus = {
	'experiment', 'test', 'trial', 'study', 'investigation', 'examine', 'analyze',
	'assess', 'evaluate', 'verify', 'validate', 'assess', 'trial and error',
	'pilot study', 'empirical', 'practical', 'hands-on', 'real-world testing',
	'field trial', 'assess the feasibility', 'proof of concept', 'explore',
	'scrutinize', 'field test', 'put to the test', 'pilot experiment',
	'validate the hypothesis', 'experimental design', 'practical examination',
	'assess the practicality', 'experiment', 'test', 'trial', 'study',
	'investigation', 'examine', 'analyze', 'assess', 'evaluate', 'verify',
	'validate', 'assess', 'trial and error', 'pilot study', 'empirical', 'practical',
	'hands-on', 'real-world testing', 'field trial', 'assess the feasibility',
	'proof of concept', 'explore', 'scrutinize', 'field test', 'put to the test',
	'pilot experiment', 'validate the hypothesis', 'experimental design',
	'practical examination', 'assess the practicality'
	}

Evaluation Dictionary	evaluation_of_results_corpus = { 'analyze', 'interpret', 'evaluate', 'assess', 'scrutinize', 'appraise', 'judge', 'examine', 'review', 'critique', 'validate', 'verify', 'measure', 'quantify', 'compare', 'contrast', 'test', 'assess the outcomes', 'draw conclusions', 'weigh',

	'gauge', 'determine', 'appraise the results', 'interpret the findings', 'evaluate the data', 'scrutinize the evidence', 'assess the impact', 'measure the success', 'analyze the performance', 'examine the results', 'analyze', 'interpret', 'evaluate', 'assess', 'scrutinize', 'appraise', 'judge', 'examine', 'review', 'critique', 'validate', 'verify', 'measure', 'quantify', 'compare', 'contrast', 'test', 'assess the outcomes', 'draw conclusions', 'weigh', 'gauge', 'determine', 'appraise the results', 'interpret the findings', 'evaluate the data', 'scrutinize the evidence', 'assess the impact', 'measure the success', 'analyze the performance', 'examine the results' }

We also used a prompt that was similar to the prompt in Table A3 in which we provided a small amount of guidance for the definition of each dimension of scientific intensity. This prompt and its resulting dictionaries in Table B6.

Table B6. Dictionary 6	
Prompt	I am performing NLP text analysis on a set of documents to examine the extent to which they reflect scientific reasoning. Please generate four dictionaries or corpuses of words that I can use to assess the extent to which the documents reflect four different components of scientific reasoning: (1) scientific theorizing, especially the use of causal logic, problem-solution formulation, and theory development; (2) hypothesis development, especially articulating assumptions, alignment with theory, and measurability and testability; (3) experimentation or testing of the hypothesis, especially the extent to which the tests are valid, representative, and rigorous; and (4) the evaluation of the results of the experiment, especially the extent to which the evaluation is systematic, based on data, and leads to better understanding.

Theory Dictionary	# Scientific Theorizing
	causal_logic = ["cause", "effect", "correlation", "influence", "relationship", "impact"]
	problem_solution_formulation = ["problem", "issue", "solution", "resolution", "address", "tackle"]
	theory_development = ["hypothesis", "model", "framework", "construct", "conceptualize", "formulate"]

Hypothesis Dictionary	# Hypothesis Development
	articulating_assumptions = ["assumption", "premise", "presupposition", "conjecture", "postulate", "supposition"]
	alignment_with_theory = ["align", "conform", "consistent", "congruent", "match", "harmonize"]
	measurability_testability = ["measurable", "quantifiable", "testable", "assessable", "observable", "verifiable"]
	# Experimentation or Testing of the Hypothesis
	valid_tests = ["validity", "reliable", "trustworthy", "sound", "credible", "authentic"]
	representative_tests = ["representative", "typical", "characteristic", "indicative", "illustrative", "inclusive"]
	rigorous_testing = ["rigorous", "thorough", "meticulous", "systematic", "precise", "exhaustive"]

Testing Dictionary	# Evaluation of Results
	systematic_evaluation = ["systematic", "methodical", "organized", "structured", "ordered", "disciplined"]
	data_based_evaluation = ["data-driven", "evidence-based", "empirical", "factual", "statistical", "quantitative"]
	enhanced_understanding = ["insight", "comprehension", "interpretation", "discernment", "clarity", "understanding"]

Evaluation Dictionary	# Evaluation of Results
	systematic_evaluation = ["systematic", "methodical", "organized", "structured", "ordered", "disciplined"]
	data_based_evaluation = ["data-driven", "evidence-based", "empirical", "factual", "statistical", "quantitative"]
	enhanced_understanding = ["insight", "comprehension", "interpretation", "discernment", "clarity", "understanding"]

We then took Dictionaries 2 and 3 above and asked ChatGPT to generate those corpuses again, but this time we added "These dictionaries should utilize laymen's vocabulary" to the end of the prompt. The dictionaries created by these prompts are shown in Table B7.

Table B7. Dictionary 7	
Prompt	… "These dictionaries should utilize layperson's vocabulary."
Theory Dictionary	# Scientific Theorizing "theorizing": ["cause", "effect", "reason", "because", "therefore", "leads to", "results in", "problem", "solution", "theory", "hypothesis", "explanation", "model", "framework", "predict", "assumption", "concept", "principle", "idea", "mechanism"]
Hypothesis Dictionary	# Hypothesis Development "hypothesis_development": ["Hypothesis", "Prediction", "Assumption", "Proposition", "Conjecture", "Postulate", "Guess", "Theory", "Presumption", "Speculation", "Anticipation", "Expectation", "Projection", "Forecast", "Scenario", "Supposition", "Belief", "Estimation", "Proposal", "Premise", "Idea", "Insight", "Hypothetical", "Thesis", "Notion", "Suggestion", "Supposition", "Anticipate", "Envision", "Forecast", "Imagine", "Postulate", "Predict", "Project", "Suggest", "Theorize", "Suppose", "Consider", "Assume", "Formulate"]
Testing Dictionary	# Evaluation of Results "experimentation_testing": ["Experiment", "Test", "Trial", "Investigation", "Analysis", "Examination", "Study", "Research", "Probe", "Evaluation", "Assessment", "Measure", "Procedure", "Method", "Protocol", "Attempt", "Verification", "Validation", "Appraisal", "Review", "Observation", "Inquiry", "Survey", "Experimentation", "Sample", "Data", "Results", "Conclusion", "Observation", "Fieldwork", "Experimentation", "Examination", "Assessment", "Measurement", "Evaluation", "Test", "Analysis", "Observation", "Experiment", "Investigation"]
Evaluation Dictionary	# Evaluation of Results "evaluation_results": ["Result", "Outcome", "Conclusion", "Analysis", "Interpretation", "Summary", "Finding", "Judgment", "Assessment", "Appraisal", "Review", "Evaluation", "Examination", "Insight", "Understanding", "Reflection", "Decision", "Consideration", "Determination", "Inference", "Appraisal", "Rating", "Critique", "Feedback", "Comment", "Report", "Summary", "Synthesis", "Conclusion", "Deduction", "Opinion", "Thought", "Criticism", "Evaluation", "Estimate", "Recommendation", "Perception", "Review", "Conclusion", "Verdict"]

After generating several options for independent dictionaries with ChatGPT, we utilized TF-IDF analysis to vectorize each dictionary as well as the corpus of interviews, then used cosine similarities to measure the distance between the dictionaries and the interviews. Cosine values are in the range of (0,1)

where higher values indicate greater levels of similarity. Thus, dictionaries with higher values correspond better with our corpus, while dictionaries with lower values do not correspond as well. Table B8 lists the average cosine similarity scores by dictionary and by dimension. As this table shows, our best performing non-lay dictionary was Dictionary 2 for the Hypothesis, Testing, and Evaluation components. The highest performing non-lay dictionary for the Theory component was Dictionary 3. The lay dictionary, Dictionary 7, performed better in all measures besides "hypothesis." We ultimately decided to use the lay dictionary for our reported regression analyses as we believe it is the most representative for the corpus as a whole. Table B8 reports the main results with the best-fit non-lay dictionaries, showing similar results to those of the lay dictionaries.[3]

Table B8. Average TF-IDF Cosine Similarity Scores

Component	Dictionary 1	Dictionary 2	Dictionary 3	Dictionary 4	Dictionary 5	Dictionary 6	Dictionary 7
Theory	0.061670	0.098014	0.195403	0.092466	0.087746	0.164624	0.208467
Hypothesis	0.019720	0.135476	0.122127	0.099248	0.051347	0.039334	0.064201
Testing	0.036890	0.189454	0.147408	0.134047	0.132782	0.066922	0.244354
Evaluation	0.012584	0.153505	0.067277	0.084751	0.072903	0.099474	0.160949

[3] We ran the same analyses with the other dictionaries and the results were consistent, so we do not report them. This process is similar to, albeit less sophisticated than, Carlson & Burbano's (2024) PVE approach.

TABLE B9. Impact of Treatment on Scientific Intensity Scores (best-fit non-lay dictionaries)

VARIABLES	(1) Scientific Intensity Human	(2) Scientific Intensity Machine	(3) Theorization Human	(4) Theorization Machine	(5) Experimentation Human	(6) Experimentation Machine
intervention	0.394***	0.031***	0.437***	0.045***	0.350***	0.018**
	(0.001)	(0.000)	(0.005)	(0.000)	(0.000)	(0.012)
Constant	1.927***	0.149***	2.316***	0.135***	1.538***	0.164***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	261	257	261	257	261	257
R-squared	0.046	0.066	0.039	0.104	0.043	0.013
Dummies for mentors	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	-	-	-	-	-	-
Clustered Errors	Intervention Mentor	Intervention Mentor	Intervention Mentor	Intervention Mentor	Intervention Mentor	Intervention Mentor

Table B10. Associations Between Frequency of Pivots (Conditional on Pivoting) and Scores (best-fit non-lay dictionaries)

VARIABLES	(1) Exactly one pivot (radical) (conditional on at least one radical pivot) (Human)	(2) Exactly one pivot (radical) (conditional on at least one radical pivot) (Machine)	(3) Exactly one pivot (any) (conditional on at least one pivot) (Human)	(4) Exactly one pivot (any) (conditional on at least one pivot) (Machine)
Theorization	-0.438***	-2.929*	-0.100**	0.329
	(0.009)	(0.097)	(0.033)	(0.602)
Experimentation	-0.438**	-3.332	-0.098	0.546
	(0.010)	(0.100)	(0.122)	(0.374)
Theorization X Experimentation	0.146***	17.409*	0.014	-2.822
	(0.006)	(0.068)	(0.405)	(0.166)
Constant	1.750***	1.028**	0.728***	0.268
	(0.001)	(0.013)	(0.000)	(0.122)
Observations	74	74	174	174
R-squared	0.173	0.111	0.110	0.045
Dummies for mentors	Yes	Yes	Yes	Yes
Clustered Errors	Intervention Mentor	Intervention Mentor	Intervention Mentor	Intervention Mentor

Two-stage Least Squares Analysis

In line with Camuffo et al. (2024), we show we find that the treatment does not lead to pivots in a linear way. While scientific intensity has a positive and precise effect with both Any or Radical Pivots Exactly Once in both human and machine-coded measures, the relationship of scientific intensity with Any or Radical Pivots At Least Once in either measure reports larger standard errors.

Table B11. Two Stage Least Square Estimates of Probability of Pivots and Scientific Intensity, Human-coded

VARIABLES	(1) Radical Pivot (At least once)	(2) Any Pivot (At least once)	(3) Radical Pivot (Exactly once)	(4) Any Pivot (Exactly once)
Scientific intensity (Human)	0.120	0.131	0.212*	0.166**
	(0.295)	(0.277)	(0.072)	(0.014)
Constant	-0.005	0.306	-0.324	-0.164
	(0.985)	(0.242)	(0.206)	(0.257)
Observations	261	261	261	261
Centred R-squared	0.082	0.262	-0.266	-0.410
Dummies for mentors	Yes	Yes	Yes	Yes
Clustered Errors	Intervention Mentor	Intervention Mentor	Intervention Mentor	Intervention Mentor

Notes. Robust pval in parentheses. *** p<0.01, ** p<0.05, * p<0. 1

Table B12. Two Stage Least Square Estimates of Probability of Pivots and Scientific Intensity, Machine-coded

VARIABLES	(1) Radical Pivot (At least once)	(2) Any Pivot (At least once)	(3) Radical Pivot (Exactly once)	(4) Any Pivot (Exactly once)
Scientific intensity (Machine)	1.980	2.061	3.457*	2.706**
	(0.333)	(0.321)	(0.055)	(0.015)
Constant	-0.112	0.206	-0.507	-0.307
	(0.765)	(0.594)	(0.125)	(0.141)
Observations	257	257	257	257
Centred R-squared	0.055	0.078	-0.146	-0.189
Dummies for mentors	Yes	Yes	Yes	Yes
Clustered Errors	Intervention Mentor	Intervention Mentor	Intervention Mentor	Intervention Mentor

APPENDIX C

Detailed Histories on Additional Cases

Table 6 in the main paper provides a classification of all the 15 cases into categories. Here we provide the case studies that, due to space limitations, are not reported in the main manuscript.

ArtisticCare: High Theorization and High Experimentation (Treatment 2)

The entrepreneur joins the program with an idea to provide artistic (visual, music, dance, acting, etc.) babysitting services to enhance children's creativity and art appreciation and a goal of employing 85 musicians. The entrepreneur's initial description of the value proposition shows a clear articulation of the problems faced by parents and their children.

"We're offering a creative experience that especially toddlers couldn't really get in a sort of a non-classical educational manner before they get to school where they're then put into sort of very traditional ways of learning the arts. This is a way for them to get creative before that happens and to sort of find their own way. And at the same time, we offer the childcare, which a lot of parents need." (Int. 0)

The idea began with a chance encounter, but the entrepreneur moved directly into testing and researching whether greater demand might exist for it.

"I tested it. I was babysitting and it turns out the parents were really happy that I was actually teaching the children some music, which is what I've been studying for. And, and then basically it was just through customer research and talking to people and asking them, would you be interested in having this creative babysitting concept? And people were all up for it and willing to pay an extra fee for that as well." (Int. 0)

As the entrepreneur began the program, they show an increased degree of theorization, articulating underlying assumptions and creating deeper cause-effect linkages across elements of their value proposition which then became the focus of their experimentation:

"the first one (we wanted to understand) was really if parents actually found that having an added component to their childcare, if that was actually something they'd be interested in. So, it's simply that. The second one was the price point, again, as I mentioned before. So, how much people would be willing to pay for such a service. Okay. And the third thing that was interesting to find out was, in terms of safety, what people feel is the most important when somebody comes to their home to take care of their kids. So, I don't know, the gender, the age of the babysitter, the languages they speak, the education. (Int.1)

The entrepreneurs utilized a variety of experiments to test their assumptions, including surveys and A/B testing. For example:

"It was a multiple choice survey (…) they just had to choose what scenarios was most appealing to them (…) Sort of a babysitter that simply puts their child to bed and watches TV with them. Another option was one that does creative activities with them. One was that tutors them in English or math (…) We asked about what was most important to them in terms of whether the age group of the sitter was important. Whether their level of education was important. Whether the gender of the sitter was important. And then we presented them with various price points. And which ones they would be willing to pay if their sitter was offering a little extra than just putting their child to bed." (Int. 1)

In line with experimentation being theory guided, the entrepreneur interprets the results in terms of how they affect the value proposition. For instance the entrepreneur notes:

"A key thing that we found out also was that parents are very keen on having the same babysitter each time and having that child build a relationship with them rather than having someone different each time. So, that was important for us to find out. And the other thing was our price point, which we found out." (Int. 1)

The results lead the entrepreneur to realize that the market may be segmentated based on location, which, in turn, leads the entrepreneur to test whether that is true:

"Well, I guess the key information we got was price point. Depending on where people were located, they were willing to pay a different amount. So we had people in North London (…) where they were willing to pay one of the highest options we gave in the multiple choice. And other areas, people who worked in London, were not even, they just chose the lowest price point we gave. "(Int. 2.)

Over time, the experimentation process revealed which elements of their theory were supported or refuted, leading the entrepreneur to pivot to targeting working mothers and specific locations, and creating different payment options (per hour vs. monthly fee). The entrepreneur used experimentation to validate the theory while integrating novel, unexpected information that arose through the experimentation process. For example the entrepreneur notes:

"We're targeting working mothers just because they're often busier and they're the ones through our research that we found have difficulties planning both babysitting and hobbies for their children." (Int. 8)

Ecoelegance: High Theorization and High Experimentation (Control 2)

This company joins the program with an idea for offering sustainable luxury clothes and beauty products. The entrepreneur joins having already done significant work with the product, including having run diligence on the supply chain and generating initial designs. The entrepreneur's goal for the year is to better develop a plan and conceptualize the implications of the business model before launching.

"What I want to do this year is basically establish the brand in the UK and establish a team, a management team and also the financials of the company. So I've just had a big business plan that I've created. So I've kind of been trying to understand the financials forecasting and cash flow and things like that a lot better. So it would be good to know my business better in that respect. And find out what's achievable and about flat planning the business" (Int 0)

The value proposition is centered around providing a luxury brand that is organic and sustainable, something the entrepreneur believes is a desire for customers:

" They want to see that is it ethical, is it sustainable, what's it made from? And Morgan Stanley did a research last year And it said that 70% of these millennials buy consciously. So I think we're offering the conscious customer an opportunity to get quality garments that they know came from a clean supply chain." (Int. 0)

Note that the entrepreneur's beliefs are the result of past experience rather than an intentional diagnosis or development of a theory:

"I think our solution will be successful because, obviously, I've had a lot of experience in the fashion industry and the industry of skincare. I'm going to say it, for me, it just felt right in my heart and then obviously I needed proof of concept within the brand and financially. So I was when we actually started trading, that's what confirmed it to me that we had such an amazing response that I believe, you know what, it's not just myself that believes it's a great brand, that our customers think that too." (Int. 0)

The entrepreneur proceeded to test and evaluate the overall value proposition. They tested customer reactions to a pre-determined set of product features more than their assumptions or specific cause-effect linkages between the customer problem and solution:

"So we've done polls, we've done statistics, we've been out on Carnaby Street in London So we've been asking customers obviously for their feedback …So we tried to have unbiased data And obviously we only did 100 people that day so we'd need to do more data for it to be more accurate". (Int.1)

With time, the entrepreneur did move to testing their theoretical assumptions. This led to a new conception of their target customer:

"The thing I wanted to understand from them is what they want, what's important to them, and…how much they're willing to pay for that…asking them to rate, you know, what we're speaking about (…) I find the customers are basically people up to over 30 years old with a lot of disposable income, they want luxury, they're sustainable, they're organic and it's a constant customer who wants to store it." (Int. 2)

These new insights lead the entrepreneur to pivot their market positioning and advertising efforts:

"what we've done is actually we've revalued our products and the way it sits in the market. We've actually increased our prices by about 40%. … because we realised that our free money that we're giving to the market is something very, very niche and that actually our prices were making us actually try and compete with non-niche brands. (Int. 2)

DataPulse: High Theorization and Low Experimentation (Treatment 2)

This company joins the program with the idea of providing data analytics services to business customers and faces a problem of revenue stability. The entrepreneurs identify and articulate a clear customer problem, although they do not explain how they arrived at that understanding:

"We work with customers to try and help them see where they can use data to drive change and transformation in their business. Usually, the issue they're facing is that they have competitors or changes in the market that's using data or data-driven automation to be more efficient, more effective or to offer completely new revenue streams." (Int. 0)

This leads to a value proposition that clearly links the problem with the proposed solution:

"So, the way we work with customers is we look at what their goals are. (…)We work with them to very rapidly build out applications that are able to take this data and turn it into operational insight that allow them to do things that are quicker or different things or to help them unlock new areas. (…)So, by helping them identify where they should focus their time for the greatest yield. (Int. 0)

The company's goal in the program is to change their revenue model to make earnings more consistent:

"At the minute, we get caught in these ever so slight boom and bust cycles where we get very, very busy and we all get flat out and we start generating good revenues and then we find that we go through a very fallow period because we've been so busy delivering work, So, we're trying to move our business to a strategic shift to have more of a relationship with our customers where we're more embedded... From a selfish perspective, it provides us with recurring revenues (…) so it gives us more time for innovation and also to smooth out our revenue cycles." (Int. 0)

After the training begins, the entrepreneur shows a more advanced level of theorization. In addition to recognizing that their challenge concerns the revenue stability, the entrepreneur is now able to identify its root cause (i.e. different customers groups with different data analytical needs and associated willingness to pay). Based on this diagnosis, the entrepreneur identifies the selection of a focused customer segment as the natural solution:

"Each of our customer segments have slightly different needs. So, we have one customer segment which is smaller businesses, and then we have a customer segment which is a much larger business. So, we had a lot of feedback on that. You know, each of our customer segments have slightly different needs. So, we have one customer segment which is smaller businesses, (…). They just want somebody to help them out every now and again, to get them moving through small bits of work. We have (…) enterprise customers, (…) and they want people to just sit there and be part of their team for six or twelve months. (…) And then we've got people in the middle who are

innovators, who want to do things in a more frameworked way, using Agile or Scrum approaches. So, we've got three separate segments, three separate solutions that we can all speak to, to be honest. We can do them all in the same way, but it's a slightly different message and value proposition to each. (…) (Int.1)

The entrepreneur engages in limited formal experimentation, consistently reporting not conducting any tests across the entire observation period:

" We spoke to our customers about what the value of our work was… We haven't done any proper, because the number is quite low, we haven't done any proper, you know, analysis or anything like that. We can do it, you know, we can eyeball it, to be honest" (Int. 1)
"Yeah, so we haven't done it yet because we've been too busy... (Int. 4)

The theory of value is refined through the internal analysis of the feedback and internal data they received through the regular operation of their business, balancing that with the preferences of the internal team. Most of the decision making is based on cognitive analysis and theorization rather than adjusting to customer feedback:

"So, the first thing we did was to benchmark where the business was in terms of our customer journey. Actually, what were we doing to generate awareness and trust? How were we converting people? What was our customer loyalty like? So, we did an initial bit of analysis where we benchmarked the quantitative side of the business. And that exposed a few things that made us question what we were doing. Our revenues weren't quite what we'd expected and our margins weren't quite what they'd expected. So, we had to try and understand why. So, we looked at our customer journey to understand where those blockers were and where those gaps were" (Int. 2)

As a result of the entrepreneur's theorization, they pivot to focusing on a specific customer segment and offering the service through a partner who lacks analytic skills. The entrepreneurs had received feedback about other kinds of services they would like, and the entrepreneurs turned that into a theory of partnership channels. Importantly, they did not move into testing that theory. Rather, they moved straight into implementation:

"We started adding a partner channel...we've been working with some partners in complementary industries, but who wouldn't have data and analysis skills like market research, HR, procurement. And we've developed, I guess, proto-partnerships in those areas with companies that we know and like... they can use our services to help their customers. " (Int. 2)

The pivot was not driven by experimentation, but was consistent with the theorization of the problem:

"The problem we're trying to solve is companies find it hard to make accurate or timely decisions because they lack the right kind of data or information...The evidence of the problem is that there's a widely published statistic that 80% of business intelligence projects are failures. And one of the key issues behind those failures is the lack of linked business processes and the lack of ability to build these applications correctly... So we help them try and know what the goals are and build the applications, give them the right information at the right time." (Int. 2).

TalentBridge: High Theorization and Low Experimentation (Control 1)

In TalentBridge (HLC1), the entrepreneur joined the program with the business idea of providing psychometric assessments and skill and career development to individuals and matching them to firms in the financial sector. The entrepreneur envisioned a hybrid organization that supported philanthropic endeavors for refugees in Middle East by leveraging for-profit operations in the UK. The business idea and organizational form were both informed by past experiences, as this entrepreneur reported having lived in

both regions for a cross-cultural understanding and a long career in the financial sector. These experiences enabled the entrepreneur to articulate the component problems (increasing GDP per capita, low skill development, insufficient banking infrastructure) and link these to the envisioned solution (providing hard currency, upskilling, upgrading banking operations), at the very onset:

"So there are three problems being solved...So there's a huge scope to improve the GDP and the GDP per capita of [x region]. So hard currency, this service was involved bringing hard currency into the country...Two is upskilling the local population...in bringing these activities on a large scale into [x region], we would need to train....And then the third problem would be for the banks in terms of upgrading their operations and activities. So international standards." Int. 0

Armed with these strong preformed beliefs, the entrepreneur engaged in minimal experimentation and proceeded directly to strategy implementation. Even when reporting conducting interviews with bankers, follow-on questions revealed that these were motivated by networking and resource acquisition rather than learning *("I've done interviews but testing per se, I really need to do some more work on the plan", Int. 2)*

The pivot emerged after failed implementation efforts, causing the entrepreneur to abandon international engagement and identify university students as an alternative target market:

in [x-region]...some of those low income workers would be university students …take a component of that and to start with I'm going to go to universities here in the UK...and offer some of the training that I would have provided ...in [x-region]. (Int. 7)

MaritimeEdge: High Theorization and Low Experimentation (Control 2)

The entrepreneur begins with the idea of providing technical consultancy in commercial maritime businesses and doing asset optimization and risk management. The entrepreneur clearly articulates the value proposition's core feature and the cause-effect linkages from the beginning. For example:

"Key customers we have, have a presence within the maritime sector. … our proposition is, we can increase their productivity, we can reduce that wastage, then … how we could do that is either by buddying up or partnering them with a business which complements their service, and secondly, is to develop some new areas of business which can work better for that company, which currently is either restricted and they're not looking in that direction." (Int.0)
"all my clients are shipping focused, so they're within shipping (…)The problems they have is, they want asset optimisations, they want to pay as little as possible to move a cargo from A to B, and if you're a ship owner, you want to get as much as revenue as much as possible. So what we do out here is, we try and optimise both the sides, either by finding the right partnership or a cargo owner, so they get a good deal on the freight, and on the ship-owning sides. (…)Most important for a business like ours is to convert an enquiry into a revenue stream." (Int. 3)

Similarly to the case of other entrepreneurs in the control group, the entrepreneur's theorization is based on the entrepreneur's prior experience:

"I worked from a SME right up to a large corporate, where I single-handedly built up a team and headed the desk for seven years. So I understand the last decade, from my network and my own experience, that engaged workforce is quite hard to find in the large corporates, and there's not been nurturing of any good talent pool" (Int. 0)

The entrepreneur does not engage into experimentation. The entrepreneur show resistance towards it, citing personal preferences against it, lack of time to do it, and a lack of need to do it because of their extensive industry experience. For instance the entrepreneurs notes:

"We've tried and tested it during our time in the trade… I have a little more than two decades of shipping expertise. So, I do... I have worked with a wide array of businesses, and I have a network of contacts which assist me in identifying these opportunities." (Int. 1)

The entrepreneur ultimately pivots away from consulting to brokering. The pivot emerges from the implementation efforts of their theoretical diagnosis, not the testing of the assumptions or linkages underlying the theory:

"We haven't changed our customer segment, but (since the last call) we divide our work into two silos…One is more time orientated… it's a limited amount of time for a closure, so when a ship is ready to find employment, to find a cargo, or a cargo is there to move, and that needs to find a vessel, so let's call it the middleman or a shipbroker's job ... that's one silo; another one is more of a consultancy-oriented business (…) we have spent more time focusing on the shipbroking aspect of it, which is very time oriented and which requires less effort. It's more risky, but the revenue stream is available fairly quickly. We've not reduced the consultancy work, but we're not taking on more projects now" (Int. 3)

MindfulHub: Low Theorization and High Experimentation (Treatment 2)

In LHT2, the entrepreneur has the idea to create a knowledge platform to improve mental well-being as an alternative to social media.

"So we thought about an alternative to the social media craze. We want to come up with a platform that will help people develop knowledge, understanding on different subjects that they might be interested in. So we're thinking of actually we are working now on a platform, an online website. In the second phase we want to develop an app and our third goal is to invent an institute, a location in London where the mental wealth would be working. So the problem that we are going to solve for people is good time management, using time in effective ways to develop knowledge instead of scrolling social media like Instagram or Facebook." (Int.0)

The entrepreneur shows a limited degree of theorization. The entrepreneur defines the value proposition largely as a function of the solution idea. The theorization does not provide clarity as to the way in which the solution will address the problem or the cause and effect linkages that will govern the problem-solution relationship. In addition, the entrepreneur does not provide insights about the platform monetization, despite the entrepreneur specifying that profitability is one intended outcome ("*we are trying to figure out which one is more profitable or easier to launch at this moment" (Int.1).* The entrepreneur mentions that the platform will be deprived of marketing but does not clarify how the platform will achieve profitability.

"So we want to develop a social network platform which is deprived of marketing and soft power messages that influence people's behaviors and thinking… we are providing valuable knowledge from experts in an attractive way for users who want to get rid of their social networking consumption, over consumption" (Int. 1)

The entrepreneur displays instead a high degree of experimentation by conducting tests consistently throughout the observation period, and, despite the claim that it verifies that the problem exists, the entrepreneur did not articulate what the problem was. The majority of these tests are conducted via a partnership with an NGO in a different country, which is mentioned in various interviews:

"I've been working with this NGO on designing a test, to run tests and to test some of our hypotheses (…) Because our focus group was only for children aged 15 to 17, they also ran some additional tests for students to age 20. (…)So, all the people that we connected and we surveyed were around the age of 20 to 35. And they admitted that they were feeling upset with the content that they saw on social media and that they felt that it created an image that is not achievable for themAnd that they want to develop their knowledge in areas that they didn't have. …I would say that it seems that our hypotheses are correct." (Int. 1)

In this case, the limited degree of theorization creates challenges in the interpretation of feedback collected:

"we have doubts about our product itself. I mean, so how should we design it? How should we serve that? Since we haven't seen anything like that anywhere, so we don't know if it's just irrelevant or if it's something that's only going to start growing in other places or that we will see very shortly from other suppliers as well…So basically, if we get subscribers, if we get attention, if we get followers on our social media, and if some of those targets are met, well, what does it say about our product or what does it say about our thinking about our business?" (Int. 1)

The entrepreneur pivots toward a more targeted customer segment, but there is no clear theoretical rationale connecting it to the value proposition. Rather the pivot is reactive to the tests results obtained from the NGO partner, which reveal that the problem is relevant also for people in their 20s.

"initially we didn't think about our customers specifically. And we wanted to address that to everybody. And now we know we need to target that more.... Yes, so that's people living in European countries, in big cities, at the capitals. And it's people aged around 20 to 35. And, yes. What else can I say is that they are very fragile about the impact of social media. Negative aspects of social media" (Int. 1)

The entrepreneur then engages in a series of additional pivots based on the survey results obtained from the NGO partner. These pivots include shifting from providing general content to young people to providing general content to families with children, then from providing general content to families with children to designing an app for providing advice and expertise with parenting to young families with young children. After this series of pivots, the entrepreneur begins to articulate a customer problem and the cause-effect linkages between it and their envisioned solution, although it is still fuzzy why the envisioned solution is the solution of choice outside of the fact that it is the solution they began with.

"And I think I realized what has been my biggest struggle with the previous value proposition that I had. Basically, I was trying to target all the people… And now we are, I think we are at the stage where we want to only target families. With children. And this is a good place to start (Int. 4)

SmartStock: Low Theorization and High Experimentation (Control 1)

In LHC1, two entrepreneurs were developing intelligent food packaging for automatic reordering by food retailers and grocers. They began the program with the main goal of finishing the product to test it with customers.

"We are an IoT based platform. Our goal is to create intelligent smart containers that allow retailers and some specialist grocers to increase their sales by enabling their customers to reorder their items…Our customers is increasing sales by automating reorder." (Int. 0)
"So the first [goal] is to finish, get the product done and have it field tested with customers." (Int. 1)

When asked about their value proposition, they remained focused on the solution itself without a clear articulation of the problems their prospective customers faced or the cause-effect linkages between the perceived problem and the proposed solution.

"The value proposition is we create intelligent devices that automate reordering. What that means is that we provide a physical device that allows someone to automatically reorder in the right quantity. The problem that we're trying to essentially solve is to make sure

that the ordering that happens or the reorders that happen are done at the right time, the right quantity to reduce wastage and reduce the total cost of inventory." (Int. 1)

When the RA asked why the customer's problem exists, the entrepreneur referred to the existing solution as being insufficient rather than describing the underlying causes of the perceived problem, whatever it was.

"So the problem exists because of a multitude of reasons. Because the way most people track inventory or count inventory for themselves or for large organizations is done manually. And it is weather-prone and it has no intelligence into it. … By doing this particular model, we are trying to change where human intervention or human activities are not the one that determines inventory levels. It's done automatically." (Int. 1)

The entrepreneurs tested their idea consistently throughout the program, demonstrating prototypes and gathering feedback from prospective customers.

"So our tests were predominantly in...demonstrating the product...and understand customer reactions to it. And that test was enormously helpful in determining what are the things that will work and what will not work...What we essentially asked them, does this product make sense? Where do you think they will find value of using this product?...What are the pricing issues that they see when selling this product? And, what kind of information and integration do we need if the product like this were to be sold? (…) So I think conclusions from the test are that there is a fair amount of user adoption areas that we need to focus on. Two is the pricing structure has to be such that there is lower risk for the customers to try because we are a new company. (Int. 1)

Despite consistent testing, there is no change in the value proposition and no evident improvement in the entrepreneurs' understanding of the customer problem. When asked why they believe that their solution will solve the customer problem, the entrepreneurs responded:

"Well, the problem still exists. Consumption tracking and food tracking is still an issue. And all the conversations that we are having with prospects are leading to field trials and experiments. They are willing to try it out. So that's giving us a positive sign that it's something that works." (Int. 5)
"We haven't changed the value proposition ... It's still the same." (Int. 5)

The entrepreneurs eventually pivot from a B2C segment to a B2B segment.

"We are more from a customer segment has changed, therefore the value proposition has moved slightly. So, we are now doing a B2B scenario where we are focusing on just inventory tracking and not necessarily looking at nutrition tracking or reorders." (Int. 6)

The pivot is about a change in customer segment and occurs based on feedback obtained while collecting evidence.

"In the last few weeks we've been talking to a few experts. So, we've got insights on how we should approach it and a few of our conversations with potential investors and customers have, and experts in that industry, have led us to believe that it's easier to sell in a B2B scenario than a B2C. And therefore we've moved in that direction." (Int. 6)

Interestingly, this is the first time the entrepreneurs articulate an actual customer problem and how their idea might address it. However, they still lack an underlying conception of the cause-effect linkages, and such a conception is not driving their evidence gathering.

"So, the conversations that we are having seem to indicate that cost optimization on the inventory side is a big issue (…)
this changed value proposition would mean that we would be able to solve other issues, such as inventory shrinkage, which is essentially theft or losses during transportation. So, we may be able to do some of it, but we're still really not clear on how that's going to pan out. Okay. But that will come to know only when we run through some field trials." (Int. 6)

By the end of the sample period, the entrepreneurs have continued to run tests and were better able to articulate basic cause-effect linkages.

"We've got what we do is we allow our customers to reduce food wastage and especially shrinkage and pilferage. Those are the things that we do. And that's the value proposition that we offer to our customers." (Int. 8)
"The target customers are predominantly food retail businesses, which are coffee shops and restaurants. And the problem that they face is unable to track how much of food is being wasted and where. So we offer a solution that allows them to manage that in real time… our product gives them real-time information, which normally is not available because most of the information is done manually. And information is outdated by the time it's entered. So this prevents any sort of manual errors and any sort of overheads associated with tracking information." (Int. 8)

However, instead of testing this rationale, the entrepreneurs continue to push forward with tests about the solution itself.

"We have data that is being, when the system runs, we gather data. And that data is what is supporting our value proposition. So we present data saying this is what, when we ran our field, you know, trials, this is what we discovered… So we talk in terms of data and analytics around that. And that's the way we try to understand how better we can sell our product." (Int. 8)

GeoEduConnect: Low Theorization and Low Experimentation (Treatment 1)

In GeoEduConnect (LLT1), the entrepreneur joined the program to pursue the business idea of offering geoscience information and education. In follow-up interviews, the entrepreneur identified how the business intended to promote awareness, and also the potential need for customization for different types of customers:

"it's an attractive way of delivering this material which is basically geological knowledge, information, guidance. So in terms of the way it's packaged as geoscience limericks or as card material, it's got quite a lot of appeal in that respect. …different things apply to the different types of customers. If it's say like university groups or school groups...they have different considerations in terms of where they are... Because the data kind of covers a wide range..they have to be specific to their age groups." (Int. 1)

However, there was no subsequent evolution in the theorization during the training period: the entrepreneur did not delve into what the differences are, and how and why these customer segments and their needs may link to potential attributes of the envisioned solutions. The entrepreneur also did not identify assumptions or cause-effect linkages in other elements of the business model in terms of value creation and capture, but continued to descript the value proposition in terms of product features:

"my product is actually about developing geo science information and engagement tools" (Int. 4)

Experimentation was also very limited, and resulted in this entrepreneur choosing not to respond to the calls for Interviews 2 and 3, as noted in their Interview 4:

[Reason for missing calls] I basically wasn't in a position to continue some of the elements...the various testing. ...I guess I did write out a thing ... I guess my tests haven't been as rigorous. ...No, I can't really speak to that, I have got things written down, but my results were inconclusive as of yet. So, I guess it's just really at an ongoing stage really, because I don't have them." (Int. 4)

The entrepreneur reported a pivot (arising because of attendance in professional conferences) in their target customer group from secondary schools and teachers to early- and mid-career scientists and the addition of an advocacy service:

The initial ones were like secondary schools and teachers, (…)I've participated in some symposium and conferences and I've actually seen that there's a need for advocacy as well as access information. So I've modified what my service I'm designing it to include include the advocacy as part of the programme so I'm trying to simplify urban geology access and advocacy because I'm looking at early career, mid-career, you know, scientists (Int. 4)

Through the end of the observation period, the entrepreneur did not articulate how the advocacy component was to be integrated, nor how it would be a value proposition to the new target group:

"during the course of the last few months, it's taken on board some new elements of advocacy as opposed to just awareness and access... we're very much in an environmentally conscious awareness now, sort of frame of mind in society. (…) In many respects, I've now expanded it to sort of take in the advocacy element and then move on to the advocacy level of promoting geoscience and environmental awareness and career development. (Int.8)

The entrepreneur also reported that no tests were conducted during this time.

LifeVantage: Low Theorization and Low Experimentation (Treatment 2)

In LLT2, the entrepreneur had the idea to offer curated media services for 45-70 year age groups and the goal to grow. The entrepreneur cites the main challenges of the business as growing:

"Challenge or issues right now as a business? Audience. That's one of our main challenges. And obviously finance. (Our main goal is) Growing, growing." (Int. 0)

The value proposition is highly focused on the idea. The entrepreneur has identified a target audience, but the entrepreneur does not articulate the customer problem or the key logic of value creation and capture:

"Our business is about targeting people over the age of 40 upwards. In order to give them ideas regarding travel, luxury items, beauty, whilst dining out, arts and culture. So, we cover a wide spectrum of topics that they can indulge in. So, it's something for everybody. That's how we like to call it. The problems I suppose we have with that is actually getting, what's the word? Maybe we're stretching ourselves too thin. And so therefore, we're not hitting the right kind of people" (Int. 1)

Experimentation was limited in the number of people the entrepreneur reached out to as well as the objective of the tests. For the tests they did run, such as an A/B test, they struggle to evaluate the results and interpret their meaning.

"This is something I feel would still like more work on like knowing how to do, analyzing the data…. There are some things that I really wished that I got more out of on the course … but from looking at my results just on the yeah, the front sheet, it was 26th, 26th was Wednesday …in terms of the age target, which is 45 plus, our highest rate is women 25% and men is 21% in terms of from the age bracket of 45 to 54. So that's not bad, that's actually one third. So in terms of the difference between men and women reading there's only 3-4% in it so it's very marginal or minimal I should say so that tells me that from the tests that I did we're on the right tracks." (Int. 1).

The entrepreneur eventually pivots to attempt to increase the readership percentage of men and (later on) the LGBT+ community. These pivots are less about changing direction and more about adding more directions to the current objectives. For example, rather than changing the target customer segment, the entrepreneur adds more to the target customer segment. Also, rather pivoting away from e.g. magazines to a podcast, the entrepreneur chooses to include a podcast. The ideas for these pivots came out of unstructured conversations with customers but the entrepreneur does not outline a clear logic behind the

change, whether evidence for the change is widespread or representative of the target customer, or how it would add value.

"we are trying to attract more male readers, so we are now going to be focusing on that. So we've closed the gap because our male readers were a lot less but now we seem to be closing the gap on that. So that has changed." (Int. 1).
"Target audience is males and male and female and the LBGT community for people over the age of 45 plus. There wasn't very much in the way for them as they were being excluded from the media, so to speak, in fashion and retail and things such as that. (Int. 6)

MatchWrite: Low Theorization and Low Experimentation (Control 1)

In LLC1, the entrepreneur is developing an editorial consultancy and editorial development to writers of fiction and non-fiction with a bespoke service whereby prospective clients are match-made with the best editor for their project. The description of the value proposition focuses the service itself, its operational challenges and details, but it lacks insight into the customer problem the entrepreneur is solving. Despite stated goals of wanting to grow, the entrepreneur does not discuss the mechanisms underlying the business model and how the growth might occur:

"I have a group of associate editors (…) my goal is at the moment, I'm getting one or two clients per month and I'd like to grow that to … I'd like that to be eight a month. the key benefit is obviously my expertise in the editorial business, which I've developed over the last 15 years. ". (Int. 0)

The entrepreneur eventually identifies a particular customer group, but it is unclear why that specific segment is chosen. The entrepreneur cites a personal ability to help that group but does not examine the causal linkages between the problem and potential solution. Furthermore, the direction is motivated by industry experience rather than strategic diagnosis of the situation.

"The problems it tries to address are to assist people, writers who are in early stages of their careers hoping to break through into publishing but possibly have already done courses but are not quite making it through to attract agents or publishers. It's to assist them along that path. So it's both actually providing edits that will get their books into better shape but also the contacts because I'm well connected within the agent world. It's both of those things. And the unique benefits and the benefits that I offer really is to do with the contacts and the experience" (Int.1)

The entrepreneur does not conduct any experimentation, citing a lack of time (" I don't have the time (to test). I just don't have the time" (Int. 4))

The entrepreneur decides to pivot to offering editing services in partnership with another business. It is unclear what causes this pivot outside of the already-existing relationship:

"I'm basically going into a partnership with a colleague of mine who's setting up a mentoring business for writers, and I'm going to offer the editing as part of one of her packages. It's going to be kind of mutually, we're going to mutually promote it and do lots of, yes, we're going to assist each other with it." (Int. 2)
"We've basically talked about this together quite a bit, but I was like, I think you need to develop this and I'll go in association because I just don't have the headspace to do it." (Int. 2)

The entrepreneur makes additional pivots such as changing the price and adding services. These pivots seem to occur in an opportunistic fashion without a strategic explanation, typically as a response to a request

from a prospective customer. It is unclear which customer segment these prospective customers are in, whether they fit the segment the entrepreneur had initially targeted or not.

"I think pricing is becoming a bit of an issue for me and I think that's something I need to decide. (…)I decided to up price at that point and (the client) was very happy to pay that was no problem. And then (…) I've had a repeat customer who came back and was all booked in to go but when I quoted him (…)but he decided that it was too much still because he had found somebody who was able to do it for a lot less" (Int. 3)
"I'm now just broadening the services that I offer, so it's line editing, copy editing, proofreading, and developmental editing. So, I've gone into the micro as well as the macro (…) People finding me through Google and then saying, do you do this? And me saying, no, but why aren't you doing that? And then deciding to offer it." (Int. 4)

DesignEdge: Low Theorization and Low Experimentation (Control 2)

The entrepreneur in this case provides interior design services. The description of the value proposition is terse. The entrepreneur does not identify a customer problem, underlying cause-effect linkages related to the service, or a strategic direction or purpose for the business.

"So we provide design services to a different range of clients from hotels to old people's homes. And then we also do research… If it's a developer and they want to create a new building, we'd help them design that, but in a way that will give them a better return for their investment. (Int. 0)

Similarly, the entrepreneur engages in limited experimentation. The genesis of this entrepreneur's assumptions is based on industry experience, informal encounters with industry stakeholders, and some secondary research.

"Not as formal as that, but more just conversations with different people in the industry… it's more sort of based on people's opinions rather than statistical." (Int. 0)

The entrepreneur pivots to a more specific type of design, designs that will "enhance people's well-being." This leads to a pivot in the choice of customer segment from individuals to companies.

"Yeah, so we're focusing on creating places that will enhance people's well-being. So that's quite a new thing. (…)
(The value proposition) is become very specialized around well-being, basically. (…) before, we worked a lot with individuals, so private homes, and now we're looking to work just with organizations and companies" (Int. 2)
"It's going to be mostly commercial from now on" (Int.3)

When asked about the reason for the change, the entrepreneur responds that it is based on a general conception of the industry rather than the result of experimentation or unique information.

"From industry trends or conferences, (wellbeing is) becoming more of a subject that people talk about… I suppose it's talking to people in the industry about the idea. Having wellbeing as part of what we do. And seeing it as a positive reaction in general, but not in any specific Int.s or anything." (Int. 2)