Entrepreneurial Operations Management

by

Evgeny M. Kagan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Business Administration)
in The University of Michigan
2018

Doctoral Committee:

Associate Professor Stephen G. Leider, Co-Chair
Professor William S. Lovejoy, Co-Chair
Associate Professor Tanya Rosenblat
Assistant Professor Eric Schwartz

Evgeny M. Kagan
ekagan@umich.edu
ORCID ID: 0000-0001-9990-1543

ACKNOWLEDGEMENTS

There are several people who made this work possible.

My deepest gratitude goes to my advisors, Stephen Leider and Bill Lovejoy. Steve's patience and ability to listen is limitless. As a teacher and a mentor, he has taught me more than I could ever give him credit for here. Bill has many qualities that make him a great advisor, but the ones I admire most is his relentless drive to make research useful for practice, and his incredible ability to tell a story (in words or in writing). I feel very lucky for the opportunity to learn from him.

I would also like to thank my dissertation committee members, Eric Schwartz and Tanya Rosenblat for their unique perspectives and their contribution to my dissertation. Additionally, I would like to thank Damian Beil for believing in me as a PhD applicant and to Roman Kapuscinski for the advice and support on the academic job market.

I am grateful to my Ross friends: Anyan, Yao, Tiffany, Murray and Iris; they were a wonderful companionship in this journey. I want to thank Uyanga who sparked my interest in doing research, Ruthy for being my big (academic) sister, Sushma, Alex, Shashank, Siyu, Remy and Kartheek for making the Ann Arbor times memorable, and Wenlu for putting up with me.

Finally, I want to thank my parents, Rita and Mikhail, as well as Peter; for their love and guidance.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

In the presence of tight capital, time and talent constraints, many traditional operational challenges are reinforced (and sometimes redefined) in the entrepreneurial setting. This dissertation addresses some of these challenges by examining theoretically and experimentally several problems in entrepreneurship and innovation for which the existing literature offers little guidance. The dissertation is organized into three chapters.

When tight time-to-market constraints are binding an important question in product development is how much time a development team should spend on generating new ideas and designs vs executing the idea, and who should make that decision. In the first chapter of this dissertation I develop an experimental approach to examining this question. Entrepreneurial ventures can have limited (often zero) cash inflow and limited access to capital, and so use equity ownership to compensate founders and early employees. In the second chapter I focus on the challenges of equity-based incentive design, examining the effects of contract form (equal vs non-equal equity splits) and time (upfront vs. delayed contracting) on effort and value generation in startups. In "technology-push" (relative to "demand-pull") innovation, technology teams often develop a new capability that may find voice in a wide range of industrial settings. However, the team may lack the appropriate marketing budget to explore each in great depth, or even all of them at any depth. In the third chapter I study entrepreneurial market identification, developing and testing search strategies for choosing a market for a new technology when the number of potential markets is large but the search budget is small.

Introduction

Why do some entrepreneurial teams succeed in commercializing a new technology or in solving an important problem, and some fail? Often, the environment plays a role: a favourable macroeconomic or political climate, a powerful incumbent or a new regulatory framework can propel or sink a rising startup. These are not the issues addressed in this dissertation, mainly because early-stage entrepreneurial teams can rarely do something about them. Instead, this dissertation focuses on the internal dynamics in a startup team: the planning and scheduling of new product development, the division of ownership in the team or the exploration of new avenues for technology commercialization. These decisions can be (and should be, as I will argue) managed proactively by the startup team to avoid predictable mistakes.

Despite the rise of interest in entrepreneurial ventures, there is little organized knowledge in the field of entrepreneurship that can be of use to the entrepreneurial practice. Much of the existing research (at the time when this dissertation was written) had been conducted with the goal of informing policy-making, and not entrepreneurial decision-making.

In the absence of frameworks to think about entrepreneurial problems, much of the guidance for entrepreneurs is derived from individual success stories and popular press, which frequently collect idiosyncratic experiences that are not validated by data. At the same time, with the rise of entrepreneurial education (offered by business schools, engineering and design schools, incubators and accelerators), there appears to

be a greater need for a more systematic investigation of the entrepreneurial dilemmas. The purpose of this dissertation is to take some initial steps in that direction.

Located at the technological core of the firm, the field of Operations Management is perhaps uniquely suited to address some of the entrepreneurial challenges. Indeed, entrepreneurial work is inherently cross-functional, mainly out of necessity, as a relatively small team needs to manage the various aspects of running a business. Further, entrepreneurial problems are inherently process and cost driven, again necessitated by tightly constrained resources available to early-stage companies. Both of these realities make the entrepreneurial context both interesting and familiar to Operations Management researchers.

Why, then do we need "entrepreneurial" Operations Management, and how is it different from "traditional" Operations Management? In the presence of tight capital, time and talent constraints, many classical operational challenges are reinforced (and sometimes redefined) in the entrepreneurial setting. As a result, the extent to which operations management research extrapolates seamlessly to startups is limited in some key contexts, and few studies are validated by data. This includes problems related to incentive design, job design, product development, and market research, which are explored in this dissertation.

Given the current, understudied nature of entrepreneurship, no single stream of work (including this dissertation) can fully answer even one entrepreneurial question. Rather, this dissertation highlights one aspect of managing entrepreneurial innovation processes – the role of human behavior.

The use of behavioral analysis, and of behavioral experiments in particular, is relatively new in the entrepreneurship and innovation literature. However, there are good reasons to use experimental methods to examine the success factors of entrepreneurship and innovation. In established and mature processes there is (often) a recipe for executing a task, decision support systems are in place, success drivers

are well-understood, and there are existing, well-defined indicators to evaluate performance (*Loch*, 2017). In contrast, the development and launch of novel products involves the innovative generation of new processes and the discovery of (rather than prior knowledge of) key success drivers. To understand these creative processes, one must closely examine internal behavioral dynamics in those processes. Experimental methods were designed specifically to achieve this goal.

The remainder of this dissertation examines three problems. The first chapter develops an experimental approach to examining time allocation in product development. In the second chapter I focus on the challenges of equity-based incentive design, examining the effects of contract form (equal vs non-equal equity splits) and time (upfront vs. delayed contracting) on effort and value generation in startups. In the third chapter I study entrepreneurial market identification, developing and testing search strategies for choosing a market for a new technology when the number of potential markets is large but the search budget is small.

CHAPTER I

Ideation-Execution Transition in Product Development

1.1 Introduction

A basic feature of product development is that the number of ideas being actively considered decreases as the development unfolds. Design texts and organizations involved in product development refer to this process as the idea or design funnel (*Wheelwright and Clark*, 1992; *Cooper et al.*, 1997; *Ulrich and Eppinger*, 2011). Especially for physical products the winnowing from many to few ideas is driven by the high costs of turning early ideas and sketches into tangible objects. As a design moves from rapid prototypes and appearance models to customer-ready versions vetted on production tool sets using genuine materials, material and tooling costs rise. There are also increasing time costs as the deadline nears and there is less time to recover from exploratory failures. Both of these realities prompt design teams to narrow their ideas to a few, and then most frequently to one, before proceeding into the more expensive development phases.

While most product design teams understand the importance of narrowing down and eventually committing to an idea, there is little guidance for when to transition from ideation to implementation and who should make this decision. In this paper

we design a laboratory experiment to study two open questions, unresolved in the literature: (1) How does the allocation of time to the ideation and execution phases of development affect the design performance and (2) does performance differ with the development team or the management making this allocation decision?

While those questions are relevant in most development situations our analysis focuses on product development contexts with the following characteristics: (a) there is a hard launch date; (b) there are rising costs as the development effort transits from ideation to implementation; (c) the product is subject to measurable, objective performance metrics; and (d) there is either a single designer or a single dominant decision maker on the design team. Development processes with hard launch dates, rising costs and objectively measurable performance characterize many physical engineered products in automotive component manufacturing, medical diagnostics, defense, industrial electronics and other industries.

Hard launch dates can derive from contractual obligations in business-to-business and business-to-government settings, industry trade shows or high selling seasons, all of which can impose serious penalties for missing the deadline. Excluded would be development processes without a hard launch deadline, for example a creative writer not under contract, one of the more speculative development efforts in a company's portfolio, or situations in which the firm can internally extend the time-to-market horizon without serious penalty.

Many physical engineered products will experience rising costs over the development effort as prototypes become more polished and use production-quality materials. It is not that design changes after the transition are impossible, but they are more costly. Indeed, the serious cost consequences of downstream ECOs (Engineering Change Orders) are legend in many industrial settings (e.g. *Loch and Terwiesch*, 1999; *Terwiesch and Loch*, 1999). However late changes may incur no additional cost in other settings, for example graphic design services or editing a novel, and our results

may not apply there.

Objective, measurable performance metrics are typical of engineering products that rely more on functional objectives and less on subjective aesthetically related ones, or where success or failure of a new product depends on the ability of the firm to match new offerings with poorly understood consumer tastes. Our results may not apply, for example, in the fashion or entertainment industries.

A single dominant decision maker (working alone or leading a team) is a formal characteristic of some efforts (for example, furniture companies which contract with well-known designers) and an informal characteristic of others. A dominant decision maker can arise organically within a team, or be a de facto reality in companies with a clear power hierarchy among the departments represented on the team. In these decisions are concentrated in the hands of one person rather than being shared. Our results may not apply in settings lacking this feature.

To be able to control the design progress and the resources (costs and time) consumed by the development it is common today for an organization to adopt some variant of a phase-review framework (*Krishnan and Ulrich*, 2001; *Ulrich and Eppinger*, 2011). At a high level these frameworks feature an "ideation" phase (where the general design strategy is determined), a "realization" or "execution" phase (where the idea is rendered in more accurate materials first using prototyping and later mass production tools and machines), and a "commercialization" phase (where all the remaining business aspects of product launch and ramp up are put into place, including supply chain formation, sales force training, communications and promotions, fulfillment, etc.). In this paper we study the first two phases. From a designer's perspective phase-review stages can also be viewed as stages of a creative process that begins with a design mandate and ends in the implementation of the chosen idea(s) in a final, fully functional product. These are creative processes in that only the design constraints are provided and designers can explore an open-ended landscape of unknown

potential for the best solution they can find that satisfies those constraints.

We argue that the time allocation to development phases and decision control may affect design behaviors with important consequences for design performance. We examine those effects in an experimental task that involves designing and building a physical object. Our experimental task is a physical design challenge that reproduces the four process features listed above. It has binding time constraints and it has two phases with a transition point after which there is an increased opportunity cost for expended materials. Designs are subject to an objective, measurable performance metric, and we study the behaviors of individual designers. Within the described context participants are free to pursue their own unique ideation and implementation strategies exploring an open-ended but searchable solution space in which the optimal is (and will forever be) unknown.

In our experiments exploration is essentially free in the ideation phase but is made costly in the execution phase. The total amount of time to complete the task is fixed by a binding deadline and is kept constant across all of our treatments while the relative allocation of time to ideation and execution is varied in the treatments. In three Exogenous schedule treatments the transition time is imposed externally. The designer is assigned to either an early (after 25% of the time), midpoint (after 50% of the time) or late (after 75% of the time) transition. Which of these is best is not clear: with an early transition point the designer may not have enough time to find a breakthrough idea but will have more time for polished execution. A later transition point allows more time for ideation but may jeopardize the timely realization of the chosen idea. Do you want to spend more time searching for a great idea, or executing a given idea? Or, would you prefer the compromise solution of transiting at the halfway point?

How flexible the transition point should be and who will ultimately make the transition decision is equally important. The designer or design team has richer in-

formation about the progress of the ideation task and may be in a better position to declare when to transit into higher cost development (*Bell*, 1969). Also, giving them ownership over the process could increase their sense of satisfaction, or responsibility or both (*Hackman and Oldham*, 1980; *Pasmore*, 1988). Being better informed and more motivated should have positive design consequences. Alternatively, the ideation phase may be more intrinsically enjoyable than execution which may delay the transition (*Boudreau et al.*, 2003). Or, the additional cognitive burden of deciding when to transit may detract from the energy invested in the ideation process. An exogenously imposed transition point could also serve as a concrete goal, which may have motivational benefits (*Locke and Latham*, 2002). Procrastination, lack of structure and/or cognitive load may have negative design consequences. Our fourth experimental treatment addresses the question of who should select the transition point by letting the designer rather than the experimenter choose the transition time.

Our study is the first attempt we are aware of to study the effects of different development schedules on design strategies and performance. Our contributions fall into three categories. First, to be able to study the internal creative process of generating and evaluating design alternatives we introduce a unique data-gathering method. This includes a new experimental task and a structured approach to tracking and recording design strategies while maintaining experimental control. The resulting data set is a rich collection of variables that capture not only how well individuals perform, but also what design activities they engage and what types of ideas they develop. The analysis of the design strategies and of the launched ideas consolidates our findings by explaining *why* certain development schedules induce better performance.

Second, our main experimental results are surprising given the conventional wisdom about the trade-off of experimentation (to find a good idea) versus execution (to implement the idea in functional form), which would lead one to suspect some monotonic or U-shaped performance in transition time. We find that mean perfor-

mance levels are statistically indistinguishable when the amount of time allocated to the ideation vs. execution phase is varied exogenously. There is, however a variance effect that aligns with intuition: both the probability of failure and mean performance conditional on non-failure increase with the length of the ideation phase, hence there is a risk-return tradeoff when choosing the length of the ideation time. By contrast, endogenously chosen transition points are uniformly worse than any of the exogenous times. That is, the designers perform worse when they have to make the transition decision on their own, compared to each of the exogenously imposed transition times. In additional treatments we examine several competing explanations and show that the dominant cause of improved performance is the clear punctuation of the exploratory and the delivery phases in exogenous transition regimes.

Third, our results add texture to several conventional design wisdoms. In particular, we find that (consistent with the conventional development paradigms) early build, testing and failing fast are associated with superior design performance, and that these behaviors occur less frequently when designers are given scheduling autonomy. That is, early physical experimentation is both a direct contributor to performance, and an observable manifestation of a more latent cognitive effect that can be influenced with managerial regimes. Another popular recommendation, "Quantity is Quality" features mixed results in our experiments, and is probably not uniformly true. Our results also indicate that the quality of generated ideas, the ability to select the best ideas and to implement the chosen idea in functional form can all be vehicles for success or failure.

1.2 Literature

The streams of literature that inform our first question (how long should the development phases be?) and our second question (who should make the allocation decision?) have few overlaps. In the following we will first discuss the OM literature

on the relative time allocation to the development phases and then move to broader psychology, marketing and job design work on creativity and project management.

1.2.1 Operational factors

The question of how to schedule product development phases has attracted some attention in OM. In an early empirical study *Mansfield* (1988) finds that Japanese manufacturers were able to improve new product quality without increasing development costs by allocating a significantly larger share of time and money to the implementation stages of the process compared to US firms which tend to spread resources evenly over the development stages. The subsequent literature considers several distinct forces driving the transition timing, however the high-level trade-off is often similar. Early transition to execution can result in insufficient exploration and poorer design choices. Late transition can facilitate the discovery of a better design configuration, but is costly in development and puts timely completion at risk (*Verganti*, 1999; *Biazzo*, 2009).

One of the objectives of product development is to achieve a product-market fit (*Krishnan and Ulrich*, 2001). Transitioning from ideation to execution early on may compromise the product-market fit especially when the market is not fully defined and downstream redesign is prohibitively costly. The time when design features are finalized should therefore depend on the pace at which market intelligence becomes available and on the ability of the firm to implement late design changes further downstream (*Krishnan et al.*, 1997). Later transition lets the design team follow the market more closely, but leaves little time for more incremental improvements that help reduce the production costs and increase the manufacturing yields (*Cohen et al.*, 1996; *Özer and Uncu*, 2013). Therefore, the transition to the execution phase should occur early when customers prioritize prices over quality assuming that the cost savings achieved during the later stages of development can be passed on to the

customers (*Kalyanaram and Krishnan*, 1997; *Bhattacharya et al.*, 1998).

The cited OM papers invoke plausible assumptions regarding the design effects of different transition times, but most are not validated with data and none delve into the behavioral drivers of those effects. The implicit assumption is that more time allocated to a stage will result in better execution of that stage. One of our goals is to explore the behavior of designers working under different time schedules in order to learn about the behavioral consequences of early vs. late transition, as well as of internal vs. externally imposed decision control. Holding the contextual (market and technological) factors constant we study the consequences of the timing of the transition and the operational autonomy on the design activities and the effects of these activities on design performance.

1.2.2 Job design and task structure

While the OM literature focuses on the factors exogenously determined by the firm's technological and market environment some worker-centric arguments can be found in the job design and work processes literature. In a series of studies of behavioral dynamics in individuals and teams working towards a deadline *Gersick* (1988, 1989, 1991) finds that individuals perceive the midpoint of the work period as a transformative moment and that this realization helps them to transition from initial learning and exploration to more execution-related activities. *Choo* (2014) presents evidence for a midpoint effect empirically in a study of Six Sigma project schedules: he finds a U-shaped effect of problem definition time on project duration. If these findings apply to design-related tasks, we should see halfway transitions resulting in better performance than either late or early transitions.

Regarding decision control *Ariely and Wertenbroch* (2002) find that individuals struggle to stick to self-imposed deadlines and perform better when a long task is split into equally spaced intervals with intermediate deliverables. *Dennis et al.* (1996,

1999) arrive at similar results using a business-challenge task in a laboratory setting. In the same vein, goal-setting theory (c.f. *Locke and Latham*, 2002) would predict that an exogenous transition time may function as a specific goal serving as an important motivator. If the advantages of time decomposition extend to design tasks, managers should impose the transition time exogenously upon the design team, rather than give them operational autonomy. There is some support for externally imposed time constraints from the human resource management literature. In particular, workers often prefer spending their time on tasks that "are the easiest, most familiar, or most satisfying"(*Boudreau et al.*, 2003) rather than allocating their time in a performance-maximizing way. Therefore, individuals may be unable to correctly allocate their time if one of the activities (e.g. exploration of ideas) is intrinsically more enjoyable than the other activities.

However, a larger part of the human resource literature would support a designer-determined transition time. Research in the job design literature supports the hypothesis that granting workers autonomy to make important decisions will positively affect performance (c.f. *Hackman and Oldham*, 1980; *Pasmore*, 1988), especially when the challenges workers face are relatively unpredictable, as would be the case in creative tasks (c.f. *Bell*, 1969, and references there). This finding has been reinforced in the product development context. Using structured interviews with product development executives *Sethi and Iqbal* (2008) show in a survey of R&D managers that when a phase-review process is enforced rigidly new product performance can suffer. *Maccormack et al.* (2001) conduct a survey of firms in the tech industry and find that flexible development processes are associated with better performing projects than processes in which the design team follows an uncompromising schedule of completion dates with stringent criteria.

To summarize the extant literature, OM models suggest that the optimal time allocation between ideation and execution can depend on contextual factors such

as technological pace, engineering and supplier flexibility, and market forces but is relatively silent on the internal behavioral and cognitive dynamics at play. Behavioral models in psychology and job design do not address our contextual setting directly, and offer (mixed) recommendations for task assignment in general. No single stream of research can be directly extrapolated to our experimental setting, in which we abstract away from external contextual detail and explore the internal consequences of varying ideation versus execution times and decision rights. Consequently, rather than forming ex ante hypotheses based on extant theory, we adopt a more inductive, exploratory approach to our data.

1.2.3 Experimental tasks in the literature

The psychology literature is dominated by tests of "creative production" (for example concept lists) that focus on ideation, or tests of "creative insight" (i.e. puzzles or riddles) that invoke an "aha" moment (*Sawyer*, 2012). Examples of the latter include the 9-dot problem and the candle problem (*Duncker*, 1945), both of which have a process dimension with the candle task also having a physical execution component. However, both tasks have only one (discovered) solution whereas the product development setting has an open-ended landscape of solutions each of which can be evaluated on a continuous scale.

There have been several attempts to study the invention of useful physical objects (*Finke et al.*, 1992; *Moreau and Dahl*, 2005) and new product definition decisions (*Ederer and Manso*, 2013; *Herz et al.*, 2014). None of those tasks reflect the development-specific structure with distinct phases and development costs increasing over time. Methodologically, our analysis is related to the studies by *Girotra et al.* (2010) and *Kornish and Ulrich* (2011) both of which examine the features of ideas generated in a business idea challenge and relate them to performance. While our experiment is different in that it has a physical execution component in addition to

the ideation stage, we also study the pool of all generated ideas and find that there are multiple drivers of performance with ideation, selection and implementation of the idea each accounting for some of the observed performance differences.

1.3 Experimental Design

To address the specifics of the product development setting we develop a real-effort physical task with an infinite strategy space. Our task reflects development contexts with (a) hard launch dates, (b) increasing costs to exploration, (c) objectively measurable performance metrics and (d) an individual designer or a strong team leader making design decisions.

1.3.1 Subjects and task description

118 subjects were recruited at the University of Michigan to participate in the study. The mean age of the subjects was 22.4. Approximately one half of the subjects were students with a major in social sciences and arts (including business and economics); the other half were students with a major in sciences, medicine and mathematics.[1] Subjects were paid a $5 show-up fee plus a payoff contingent on their performance in the design task. The total payoff including show-up fee ranged from $6 to $32. Participants worked individually on the following task: given 10 playing cards and 10 paper clips build a structure as tall as possible that will support as many coins as possible (up to a maximum of 16 dollar quarters).[2]

Participants were informed that the task consisted of two phases. During phase 1 all participants were given ample materials to experiment and explore. During phase

[1] Appendix A.4 presents detailed demographics data.

[2] For the exact transcript of the experimental instructions see Appendix A.4 . The full set of instructions including the description of all measures collected during the experiment is included in the online supplements (http://webuser.bus.umich.edu/ekagan/research.html). Our experimental task is a version of a challenge used at creativity competitions among high schools and colleges. See, for example, `http://www.iu13.org/images/uploads/documents/IS/PULSE/PULSE_newsletter_Feb2014.pdf`, page 4.

2 participants were given only 10 cards and 10 clips to work with. At the end of the experiment, each participant was required to present his/her structure that contained at most the 10 cards, 10 clips and 16 quarters they were given. Participants were free to use their time as they saw fit and were only constrained by the amount of materials in phase 2. In particular, they did not have to replicate the phase 1 design during phase 2.

Participants were paid based on the performance of their final design in phase 2. Performance was determined by the product of the number of coins and the construction height. The following formula was used to scale the payoffs to an average hourly rate of approximately $15:

$$\frac{\text{Monetary value of supported coins} \times \text{height of the structure in inches}}{3}$$

1.3.2 Experimental procedures

In all treatments participants were given 20 minutes in total.[3] They were randomly assigned to one of four treatments. Three treatments featured an exogenously imposed transition from the ideation phase to the execution phase while varying the shares of the time allocated to the phases. In these treatments participants were given 5 (10, 15) minutes for ideation, after which ideation materials were taken away. Then participants received the second (exactly 10 cards and 10 clips) set of materials, and were asked to build the structure that was to be submitted for performance evaluation. They then had 15 (10, 5) minutes to finish their work. In the fourth treatment we asked participants to choose their own transition time. Participants were instructed

[3]When choosing the appropriate duration of the task our objective was to impose binding time constraints, and at the same time provide enough time for exploration of the design space. To calibrate the allowed development time we ran a pilot session with 9 participants. The task duration in the pilot was 20 minutes. Each subject was able to complete the task with payoffs ranging from $2.67 to $18.67. All subjects appeared to be working throughout the duration of the task, i.e. the time deadline was binding. The pilot data are not used in the presented analysis due to minor differences in the instructions. However, including the pilot data does not affect the performance results.

to raise their hand to indicate the transition to the execution phase, after which their exploration materials were collected and the second (constrained) set of materials was distributed. We refer to this treatment as the Endogenous treatment.

We used a between-subjects design with 4 experimental sessions run in each treatment. Each participant was monitored discreetly by a camera placed behind a one-way mirror located close to the ceiling of the laboratory.[4] Throughout the experiment participants were separated from their neighbors by partition panels. Remaining time was announced every 5 minutes and a clock was projected on a large screen, visible to all participants. Upon completion of the design task we elicited subjects' risk and ambiguity attitudes using the Holt and Laury method (*Holt and Laury*, 2002) and administered the Need for Cognitive Closure survey (42-item questionnaire about uncertainty attitudes, *Webster and Kruglanski*, 1994).[5]

In section 5 we will consider three additional treatments in which transitions were also determined endogenously by the designers, but either the information provided to the designers or the mechanics of the transitions differed. The specific details of the experimental procedures for those additional treatments will be discussed later.

1.4 Experimental Results

The remainder of this paper is organized as follows. This section will investigate whether design performance and design activities vary with the transition time and with the initiator of the transition. Sections 4.1-4.3 will examine the differences at the performance level. Sections 4.4-4.6 will use the video data to study the micro-process engaged by the designers (discussion of the video-analysis methodology is postponed until section 4.4). Section 5 will examine several additional scenarios with endogenous

[4] A consent form informing the participants about the videotaping was distributed and signed before the experiment.

[5] In addition to the main task earnings participants could earn between \$1 and \$5 from the elicitation of risk and ambiguity preferences. None of the elicited risk and ambiguity attitude measures were significantly related to design performance.

transitions and section 6 will re-examine our data focusing on the relative importance of idea generation, selection and implementation.

1.4.1 Performance comparisons: Measurement

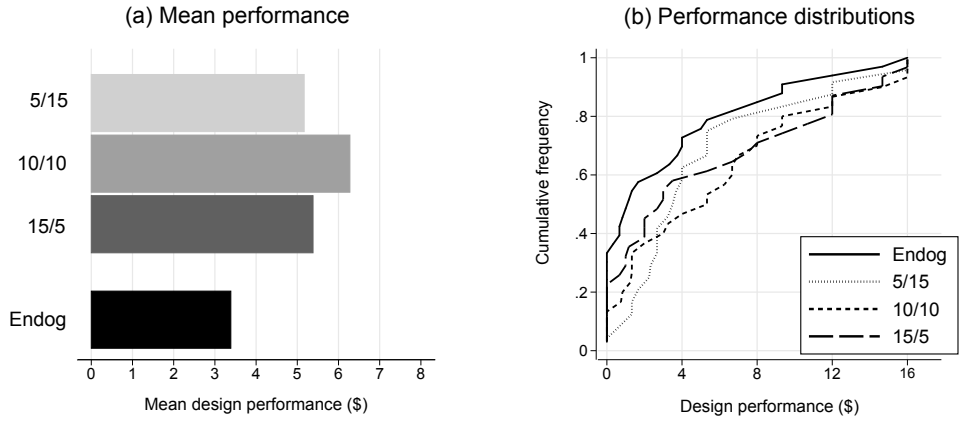
Performance is measured as the dollar payoff obtained in the design task. We begin with performance comparisons across the four treatments using two-sided non-parametric tests. We then present more precise estimates of mean performance differences obtained in OLS and Tobit regressions controlling for demographic differences and endogenously chosen transition times. We further examine the performance distributions generated by each treatment using tests of stochastic dominance, tests of equality of variances and quantile regressions. We will sometimes use the short notation 5/15, 10/10, 15/5 when referring to the three Exogenous treatments, Exog when referring to the pooled Exogenous treatment and Endog when referring to the Endogenous treatment.

1.4.2 Performance comparisons: Results

Design performance comparisons

Mean performance in each treatment is shown in panel (a) of Figure 1.1. The differences between any two of the three Exogenous treatment groups are not significant at any conventional level (Rank Sum test, lowest $p = 0.491$). In contrast, there is a significant difference of \$2.25 between the means of the pooled Exogenous and the Endogenous treatments (\$5.64 vs \$3.39, Rank Sum test, $p = 0.012$). That is, on average the design performance is improved by about 66% when a designer's schedule is changed from endogenously determined to exogenously imposed. Both the means and the medians of performance in each of the Exogenous treatments are higher than in the Endogenous treatment with the 5/15 and 10/10 treatments being significantly better (Rank Sum test, $p = 0.024$ and $p = 0.020$). The 15/5 treatment

Figure 1.1: Performance across treatments



Note. In panel (a) bars show mean treatment performance ($). Panel (b) shows within-treatment performance distribution. Three observations in the 10/10 treatment feature performance greater than $16 ($17.3, $17.3 and $24). Support values of performance distribution in the other treatments reach a maximum value of $16. For presentation purpose in panel (b) these three observations in the 10/10 treatment have been assigned a value of $16.

is not significantly different from the Endogenous treatment despite having a higher mean and median (Rank Sum test, $p = 0.135$). This is driven by the high dispersion of performance outcomes in the 15/5 treatment rather than by a smaller magnitude of the difference.

While different exogenous allocations of time to development phases do not change mean performance, they do affect the likelihood of design failure. 23% of participants in the 15/5 treatment are not able to build a viable structure as compared to 13% in the 10/10 treatment (Two-sided Proportion test, $p = 0.348$) and 4% in the 5/15 treatment ($p = 0.055$). The occurrence of failures rises monotonically with the length of the ideation phase for the Exogenous transition (Trend test, $p = 0.052$, *Cuzick*, 1985). However, at 33% the proportion of zeroes is the highest for the Endogenous transition group with the difference between Endogenous and pooled Exogenous treatments being significant at $p = 0.018$ (Proportion test). The percentages of design failures, mean performance and mean performance conditional on non-failure are summarized in Table 1.1.

Table 1.1: Summary statistics of participant performance by treatment

Treatment	% Failures	Mean performance ($)	Mean performance given non-failure ($)
Endogenous	33.33	3.39	5.08
5/15	4.17	5.17	5.39
10/10	13.33	6.28	7.24
15/5	22.58	5.38	6.95
All treatments	25.31	5.01	6.22

Notes. % Failures column shows the percentage of participants unable to present a valid structure after 20 minutes. Mean performance column shows performance measured as the dollar payoff obtained in the design task (excluding the show-up fee of $5 and payoffs from uncertainty attitudes elicitation). Mean performance given non-failure column shows mean performance of the subjects who were able to present a valid structure.

Regression results in Table 1.2 confirm the results of non-parametric tests. Participants in the Endogenous transition group perform uniformly worse than each of the Exogenous treatment groups. Columns (1) and (2) show Probit marginal effects with non-failure as the dependent variable. When the decision-maker is concerned with minimizing the risk of design failure the 5/15 and the 10/10 treatments are both significantly better than endogenous decision control. In contrast, 15/5 is not significantly different from the Endogenous treatment.[6] Columns (3) and (4) report the OLS coefficients with design performance as the dependent variable. Given a baseline performance of $3.39 obtained in the Endogenous transition treatment (Endogenous treatment is the omitted dummy variable in all regressions in Table 1.2), performance differentials range from $1.78 to $3.18 depending on the specification and the assigned Exogenous treatment.

Columns (5) and (6) report Tobit regression coefficients accounting for the clustering of performance outcomes at zero to improve the precision of the estimates and also allowing estimation of the (conditional) treatment effects for non-zero perfor-

[6]To check whether multiple hypothesis testing had a notable influence on our results we calculated Bonferroni-Holm adjusted p-values (*Holm*, 1979) for this and other important results. Multiple hypothesis adjustment has been suggested in the experimental literature to counteract potential type I errors resulting from testing the effects of multiple independent treatments on the same outcome variable (c.f. *Athey and Imbens*, 2016; *List et al.*, 2016). For additional details on the adjustment methodology and for the summary of results see Appendix A .

mance. Unconditional marginal effects range from \$2.19 ($p = 0.054$) for the 15/5 treatment to \$3.08 for the 10/10 treatment ($p = 0.010$). Conditional on non-failure the treatment effects range from \$1.57 to \$2.23 accounting for approximately 72% of the unconditional marginal effects.[7]

In sum, each of the exogenous schedules dominates the endogenously determined schedules. The treatment effects on performance can be traced in part to design failures, but these do not fully explain the results since a substantial gap remains after controlling for non-failure.

Variance effects in performance.

Especially for creative tasks the decision-maker may be interested in the right tail of the performance distribution rather than in measures of central tendency, so it is useful to examine the entire distribution of performance in each treatment. Figure 1.1b) suggests that each Exogenous treatment dominates the Endogenous treatment in the sense of First Order Stochastic Dominance (FOSD). Formally, FOSD tests (*Anderson*, 1996; *Ng et al.*, 2011) confirm the dominance in performance of the (pooled) Exogenous treatments.[8] This means that the Exogenous treatments would yield a higher expected utility for the Endogenous treatment for any decision maker with a non-decreasing utility function.

While the pooled Exogenous treatments dominate the Endogenous treatment at

[7] Age and college major help identify two subpopulations of subjects who performed significantly better than the rest: subjects who were enrolled in sciences, mathematics and engineering ($n = 52$) and older subjects (median split by age, resulting in $n = 59$). For robustness we ran all regression specifications on these subpopulations. Treatment effects are greater in magnitude relative to the full sample: unconditional average marginal effects are between 2.91 and 4.64 for the Tobit specification in column (6), p-values are between 0.013 and 0.097.

[8] *Anderson* (1996) is a non-parametric test based on splitting the combined performance data into equally spaced intervals and then comparing the number of observations in each interval between treatment groups. Using a quartile split we find that the Endogenous treatment is dominated by the pooled Exogenous treatments ($p = 0.027$). *Ng et al.* (2011) method uses quantile regression coefficients (and their asymptotic distributions) to determine whether one group has consistently higher/lower marginal effects over a range of quantiles. Using this method we are able to reject the Null of the non-dominance of either of the two distributions at $p < 0.05$.

any given quantile, the size of the performance gap depends on the transition time and the quantile range (see figure 1.1b). In particular, the performance gap between 10/10 and the Endogenous treatment remains substantial ($2-$4) at any within-group quantile. By contrast, the gap is relatively narrow ($0-$2) for the bottom 60% when comparing 15/5 and the Endogenous treatment and for the top 40% when comparing 5/15 and the Endogenous treatment. This suggests a variance effect in transition time. Indeed, the 5/15 and 15/5 treatment cdfs exhibit a single-crossing property implying that the preferred exogenous regime depends on the risk preferences of the decision-maker in question. The variance increase is confirmed by tests of equality of variances (*Levene*, 1960). The variance in the 5/15 treatment is lower than the variance in pooled 10/10 and 15/5 treatments, and also lower than in the 15/5 treatment with the difference being marginally significant ($p = 0.069$ and $p = 0.075$).[9]

In sum, while there are no differences in mean performance there are variance effects in performance within the Exogenous treatments. While a risk neutral decision-maker would be indifferent in transition time (as long as it is imposed exogenously), a risk-averse decision-maker would avoid long ideation phases while a risk-seeking decision maker would avoid short ideation phases.

Endogenously chosen transition times.

We next study *when* designers make the transition when they are given the decision rights and whether performance differs with the endogenously chosen times. The average endogenously chosen transition time is 10.74 minutes. With endogenous transitions, later transition times are associated with significantly reduced performance

[9]The following example illustrates the preference order conditional on the degree of risk aversion. Suppose a decision-maker is an expected utility maximizer characterized by the power utility function $u(x) = x^a$. Given the performance data in the Exogenous treatments, she prefers 5/15 for $a \in (0, 0.44)$, and 10/10 for $a \in [0.44, \infty)$. The least preferred exogenous allocation is 15/5 for $a \in (0, 0.85)$ and 5/15 for $a \in [0.85, \infty)$. That is, later transition is characterized by both greater upside and greater downside risk, but there are some non-linearities in the underlying data (the marginal improvement in performance given non-failure is strong as one goes from 5/15 to 10/10 but the improvement is negligible as one goes from 10/10 to 15/5).

Table 1.2: Performance comparisons across treatments.

	(1) Probit	(2) Probit	(3) OLS	(4) OLS	(5) Tobit	(6) Tobit
5/15 treatment	1.301**	1.416***	1.780	2.372*	3.039*	3.688**
	(0.513)	(0.523)	(1.217)	(1.310)	(1.679)	(1.720)
10/10 treatment	0.680*	0.794**	2.891**	3.177**	3.859**	4.181**
	(0.368)	(0.401)	(1.363)	(1.363)	(1.589)	(1.608)
15/5 treatment	0.322	0.395	1.996	2.406*	2.602	3.093*
	(0.338)	(0.361)	(1.268)	(1.311)	(1.585)	(1.598)
Constant	0.431*	0.390	3.386***	-2.866	1.985*	-4.473
	(0.227)	(0.796)	(0.783)	(3.894)	(1.125)	(3.997)
Controls	NO	YES	NO	YES	NO	YES
Observations	118	112	118	112	118	112

Notes. Probit, OLS and Tobit coefficients are reported. The omitted category is the Endogenous treatment. Dependent variable is non-failure ($\mathbb{1}_{Performance>0}$) for Probit and continuous Performance ($) for OLS and Tobit. Endogenous treatment dummy is omitted in all specifications. Controls are age, gender and Engineering major (Yes/No). The difference in the number of observations is due to six subjects not providing demographic data.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

($\rho = -0.365, p = 0.037$). That is, while performance is invariant in transition time with exogenously determined schedules, performance deteriorates in transition time for designer-determined schedules.

Exogenous transitions lead to significantly improved performance both before and after controlling for the transition times. The pooled Exogenous treatment has an average advantage of $2.77 (Tobit regression, $p < 0.01$) relative to the Endogenous treatment. After controlling for the transition times the gap is almost unchanged at $2.76. However, after adding the interaction term between the Endogenous treatment dummy and the transition time we find that the performance gap between Endog and Exog increases in transition time. The performance effect of Exogenous transition is negligible and not statistically significant when comparing performance at the 5^{th} minute ($0.48, $p = 0.793$). However, it increases in magnitude and statistical significance with later transition time reaching $2.49 at the 10^{th} minute and $3.91 at the

15^{th} minute (the effect is significant at $p < 0.05$ starting with minute 9).[10]

Summarizing the performance comparisons, in the Exogenous treatments the increase in risk of failure with longer ideation is at least partially offset by the improved performance of non-zero constructions. That is, later transition increases the risk but does not affect mean performance. In contrast, performance in the Endogenous treatment is uniformly worse than in any of the Exogenous treatments with later endogenous transitions performing worse than earlier endogenous transitions.

1.4.3 Performance comparisons: Discussion

Given that designers have heterogeneous abilities and may differ in their exploration strategy one could expect that they are in a good position to decide when to initiate the transition and start execution. We have seen the opposite, that the Endogenous transition treatment does worse than any of the Exogenous treatments even after controlling for age, major and the endogenously chosen transition times.

Endogenous treatment participants tended to fail more, garnering zero reward, but even restricted to non-failures the Exogenous scenarios are better than the Endogenous scenario. In fact increased failures explain only 1/3 of the advantage of the Exogenous treatment. The advantage of the Exogenous decision control extends to the entire distribution of performance outcomes with the Endogenous treatment being first order stochastically dominated by the combined Exogenous treatments. The negative effects of internal decision control persist for comparisons at any within-group performance percentile, so a decision maker prefers the Exogenous treatments as long as her utility function is non-decreasing in payoffs.

Can the inferior performance of the Endogenous transition group be explained by the transition times they choose? We explored this alternative by making perfor-

[10]The coefficient of the interaction between decision control and transition time is not statistically significant: $\beta = -0.560,\ p = 0.104$. Rather than focusing on the average interaction effect our analysis uses the interaction to estimate the effects of decision control holding transition time constant.

mance comparisons between Exogenous and Endogenous groups holding transition times constant and found that transition time and decision control interact. When designers have to choose transition times for themselves we see a deterioration of performance with later transition times suggesting that late transitions are at least partially responsible for the observed treatment differences. A plausible interpretation of this finding is that designers who transition late are forced into executing their design under time pressure and this hurts performance. Few of them are able to exploit a long ideation phase to find a truly exceptional (and executable) design. However, even when the time split is 50-50 the gap between the Endogenous and the Exogenous groups remains significant, so poor performance of those who choose to transition late does not explain all of the performance gap.

In contrast to the substantial differences between the Exogenous and Endogenous treatments, all of the Exogenous groups do similarly well in mean performance. However, we find that the risk of failure (zero payoff) increases in transition time, suggesting a risk-return trade-off that aligns with intuition. Longer ideation times and shorter execution times are higher risk schedules. The converse, short ideation times and long execution times have the lowest performance improvement relative to the endogenous base case conditional on non-failure. This makes intuitive sense. While it is not clear that these effects will exactly balance (so there is no statistical difference among treatments) in more general cases, we would still expect risk-averse decisions makers to prefer shorter ideation and longer execution times.

Our review of the job design and the organizational psychology literature (section 2.2) suggested three potential mechanisms that may drive poor performance in the endogenous transition regime. The first one, the idea that the intrinsic enjoyment of one of the phases may prevent an efficient endogenous allocation of time is not supported in our data. Mean performance does not vary with exogenous time allocation, so in principle any (reasonably) chosen transition time could lead to good performance.

Rather, there appears to be something about the exogeneity of the time constraint that improves design performance. Either of the two remaining mechanisms suggested in the literature (increased cognitive load in endogenous transitions and motivational effects of process goals) could drive the performance gap. The next sections will further unpack the performance advantage of exogenous transition regimes by analyzing the design activities (sections 4.4-4.6) and by examining alternative managerial regimes that keep transitions endogenous but change some aspects of the transition process (section 5).

1.4.4 Design Process: Measurement

We continue the investigation by looking at the micro-structure of the creative effort and examine what behaviors are related to improved performance and whether those behaviors differ with the time allocation and decision control. Using individual-level videos we were able to record (a) subjects' activities, i.e. their exploration and testing strategies and (b) the structural properties of the ideas they launch. The examination of the design process data helps develop intuition for what is good design practice and how the design strategies and the launched ideas differ with endogenous/exogenous decision control.

Data-gathering approach

To allow insight into the micro-structure of the creative exercise and its relationship to outcomes the video data first required an interpretive stage to go from the raw data to data amenable to statistical analysis. Qualitative research techniques were designed specifically to achieve such mappings. The body of work on qualitative methods is now extensive (c.f. *Strauss*, 1991; *Miles and Huberman*, 1994; *Maxwell*, 2012; *Saldaña*, 2011; *Yin*, 2013, and references there), and converges on "coding" as the method of choice to map unstructured inputs into more highly structured data.

A code is a symbol (letter, number, word or phrase) that reflects the content of a segment of qualitative data. Researchers derive codes either inductively (looking at the visuals and cataloging what appears to be happening) or deductively (constructing categories based on existing theory and/or the research questions being asked) or, as in our case, a combination of the two. Since we were looking at idea generation and execution, our attention was naturally focused on those and related activities. Then, once a code is derived, the compromising effects of subjectivity are reduced by having independent researchers code the videos (mapping visual inputs into code categories with time stamps), and further reduced by using multiple independent coders and looking for consistency among them.

In most cases, and in ours, deriving a usable coding scheme is a time-consuming iterative process. Each of the co-authors reviewed videos and proposed a scheme designed to capture subjects' behaviors, and then all co-authors attempted to use each scheme on a varying test set of videos in a search for agreement. After several convergence failures with alternative coding schemes we generated one based on cataloging the structural elements of an idea and reviewed the final structures generated by the subjects to assure comprehensiveness (see the electronic companion at the end of this document for our final coding scheme and figure 1.2 for examples of some structures and their codes). We then recruited and trained student coders and asked that each coder analyze each video and record the results in a data sheet. These data sheets were checked for inter-coder consistency and then used as inputs to our analysis.

The coders were unaware of the experimental results and the research questions. The coded variables were aggregated by averaging the values submitted by the coders. Each coder first worked on three training videotapes (which covered a wide range of construction strategies) to provide a sufficient level of understanding of the tracking and classification method. To ensure that coding outcomes did not interact with the treatments we assigned and randomized the order in which the videotapes were

coded. The data set was divided into 8 parts with each treatment split into 2 parts. The order in which the coders performed the coding was ABCDABCD for coder 1, BCDABCDA for coder 2 and CDABCDAB for coder 3.[11]

Recording design ideas

The main building block of our coding approach is a "design idea" or a "design strategy" which characterizes the basic appearance features of each construction launched by a designer. Each design idea was characterized by four attributes:

1. general form (P/WB/ML/Fl)
2. load bearing strategy (V/A)
3. integration of components (SEP/MP)
4. use of materials (F/T/P)

The first attribute, the construction's general form can be a "pedestal" (P), a "wall/box" (WB), a "multiple-legs" structure (ML) and a "flat stack" (Fl). A "pedestal" is characterized by a narrow load-bearing surface. The difference between WB and ML is that WB features multiple, connected (or visibly touching across a large surface) cards making up a wall. ML has several stand-alone "legs" connected only on top. Fl is a flat stack of cards piled horizontally. The second attribute of an idea is its load bearing strategy which can be vertical (V), angled (A), or both. The third attribute refers to how the construction components (coins or layers) are connected. Components can have a separating surface card between them (SEP) or a multi-purpose surface, for example when coins are placed directly on the sharp corner of a folded

[11]Our main concern in creating a randomized coding order were possible learning and fatigue effects. The total net runtime of the videotapes exceeds 40 hours and coders typically spent an additional 40 hours interpreting and filling in the coding forms. The supplementary documents at http://webuser.bus.umich.edu/ekagan/research.html provide several inter-coder reliability measures for the design strategy variables. Most of the variables show high levels of consistency.

Figure 1.2: A sample of design ideas

Notes. The images are examples of single-level and two-level structures annotated with their code.

card. The fourth attribute, use of materials can include folding (F), tearing (T) and piercing (P), or any combination of those elements. Figure 1.2 demonstrates several ideas along with their assigned codes.

Many construction ideas featured more than one layer. For such multi-layered constructions first each layer was characterized using the 4-attribute vector. Then,

if layers were identical the entire construction was characterized using the layer code (c.f. the leftmost construction in the bottom row of figure 1.2). If a construction exhibited two or more different layers the attributes of each layer were included in the code (see for example the three rightmost constructions in the bottom row of figure 1.2).

Variable definition

Assigning a descriptive code to each idea creates a clear rule that helps distinguish a new idea from a variation on an existing idea. A new idea was recorded each time any of the four elements of the idea code were changed. The change of code could either be triggered by a change of the structural properties of an existing construction or by an addition of a layer with a hitherto unused structural property. We used the number of design ideas each subject entertained before committing to a design as a measure for idea quantity.

In addition to counting the number of ideas we recorded the times when each idea was launched and when it was abandoned. Similarly, we tracked and recorded several other activities engaged by the designers. In particular, we recorded the number of coin stacking attempts, the number of construction collapses as well as the times when those events occurred. The descriptive statistics of these variables are presented in Appendix A.[12]

[12]Our list of variables initially included the number and times of variations on each idea. However, due to low consistency (correlations across coders < 0.3) those variables were discarded. We also attempted to combine measures of search behavior and testing/failures by looking at the number of ideas with at least one collapse/failure, as well as number of collapses/failures per idea. These measures showed low levels of inter-coder consistency and did not predict performance.

1.4.5 Design Process: Results

Design activities and design performance

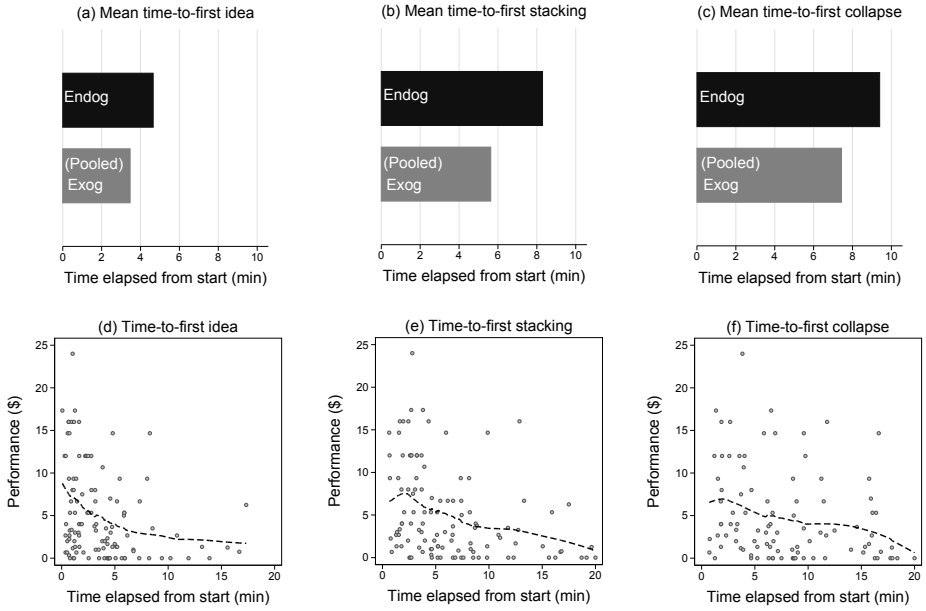
With exogenous decision control the frequency of failures and mean performance given non-failure both increased with later transitions leading to similar mean performance levels in the Exogenous treatments. The design process data are consistent with those performance results. In particular, in the Exogenous treatments most process measures exhibit only mild differences in transition time. The only significant trend can be observed for the times-to-last stacking (16:28 min in 5/15, 16:57 min in 10/10, 18:29 min in 15/5, $p < 0.01$). However the relationship between this process measure and the payoff measures is weak (Exogenous treatments: $\rho_{\text{payoff, time-to-last stacking}} = -0.22$, $p = 0.046$, $\rho_{\text{probability of failure, time-to-last stacking}} = -0.01$, $p = 0.950$).[13]

In contrast to the within-Exogenous treatment comparisons there were some striking differences in the design activities engaged in the (pooled) Exogenous and in the Endogenous treatments. We focus in particular on the differences in activity timing.[14] Figure 1.3a)-1.3c) shows that the times of launching the first idea, the first test and experiencing the first collapse each occur with a substantial delay in the Endogenous treatment (Rank Sum tests, $p = 0.017$, $p = 0.011$, $p = 0.091$), relative to the pooled Exogenous treatments. That is, designers in the Endogenous condition

[13]While the differences in the times-to-last stacking are statistically significant, the differences in the times-to-last idea are not (6:22 min in 5/15, 07:43 min in 10/10, 08:05 min in 15/5, $p = 0.156$). In the 5/15 treatment the last idea is launched after the transition to execution ($6:22 > 5:00$, one-sided $t-$test: $p = 0.088$) and in 10/10 and 15/5 treatments the last idea is launched before the transition ($07:43 < 10:00$, $p = 0.014$ and $08:05 < 15:00$, $p = 0.000$). This suggests that there is some degree of endogeneity in how much time is spent on ideation even with exogenous transitions. An alternative measure of time allocation between ideation and execution is the number of ideas explored after a successful stacking. Trying new ideas after the first successful stacking indicates that a designer is engaged in a broader exploration of the design space, as opposed to stopping and polishing an idea that works. Similar to the time-to-last idea, there was a mild trend of exploring more new ideas after a successful stacking with later transition time, however the trend was not significant (Exog treatment means: 0.49, 0.62, 0,75; Trend test: $p = 0.239$).

[14]There were treatment differences in other process variables, however those process variables failed to predict performance, so we do not discuss them here. See Appendix A for the complete list.

Figure 1.3: Activity times and design performance



Note. Figures (a), (b) and (c) show mean times to first idea, stacking and collapse by treatment. Figures (d), (e) and (f) show the relationships between those variables and performance. The dotted line indicates locally weighted scatterplot smoothing (bandwidth = 0.8).

spend more time pondering about possible design strategies or exploring the materials before launching a (recognizable) design idea. We also ran duration analysis with time-to-first stacking and time-to-first idea as dependent duration variables. Consistent with the non-parametric test results Endogenous transition is associated with longer time-to-first build and with longer time-to-first stacking. For example, for the time-to-first stacking the Endogenous treatment is associated with a delay of 2.86 minutes ($p < 0.01$).[15]

[15]Cox Proportional Hazard model was used to estimate marginal effects in the duration analysis. Additional analysis revealed that the delays in first stacking were also associated with reduced exploration measured as the number of ideas after a successful stacking ($\rho = -0.33,\ p < 0.01$) and with reduced testing intensity measured as the number of successful stackings on new ideas after a successful stacking ($\rho = -0.28,\ p < 0.01$). Similar results were obtained for the delays in time-to-first idea. Both the number of ideas after a successful stacking and the number of successful stackings on new ideas were significantly reduced in the Endogenous treatment (Rank sum tests, $p < 0.01$ and $p = 0.024$) and both were also significantly related to payoff ($\rho = 0.18,\ p = 0.051$ and $\rho = 0.19,\ p = 0.042$). That is, delayed physical ideation in the Endogenous treatment results in insufficient exploration of the design space and insufficient testing, leading to reduced performance.

Correlation analysis reveals that the tracked count variables (number of construction ideas, stackings and collapses) are not significantly related to design performance (all $\rho < 0.15$, all $p > 0.1$). However, times-to-first idea, stacking and collapse are associated with greatly improved performance, as shown in figure 1.3d)-1.3f). In particular, the ability to create a viable structure early on is associated with improved performance ($\rho = -0.308,\ p < 0.001$). Performance is also improved when the first coin stacking occurs early on ($\rho = -0.349,\ p < 0.001$) and when the first failure occurs early on ($\rho = -0.250,\ p = 0.014$). Similar results were obtained in regression analysis after controlling for individual differences.[16]

Do process variables explain treatment differences in performance?

Our results so far indicate that both decision control and delays in important activities explain large portions of performance differences. We have also seen that endogenous decision control is associated with delays in each of those activities. However, it is unclear how much of the treatment differences in performance are explained by the process delays, and how much remains unexplained. To examine the relative contribution of the process delays to the performance gap we regressed performance on both the Endogenous treatment dummy variable and the process variables. Table 1.3 shows the Tobit coefficients and the percentages of common variation in performance explained by the process variables.[17]

[16]To test for non-linearities in the relationship between the timing of ideas and performance we created variables for the number of ideas in each 2 minute interval of the 20 minute period. This alternative specification confirmed that more ideas led to better performance when those ideas were explored in the first two minutes ($p < 0.01$), whereas the number of ideas in later periods did not affect performance. Similar results were obtained for 3, 4, 5, 6 minute windows, highest $p = 0.022$. We conducted similar robustness analyses for times-to-first stacking and collapse, both of which were consistent with the presented results.

[17]The percentages were calculated as follows. We first calculated the marginal effects of the process variables and of the Endogenous treatment. We then calculated the predicted performance differences using those marginal effects and average treatment values of the process variables. This was done by taking the ratio of the predicted difference between the Endogenous and the pooled Exogenous treatment that is due to the process variable and the total predicted difference (that is due to both the treatment dummy and the process variable). For robustness we also tested several alternative specifications that included multiple process variables, discretized process variables and

Table 1.3: Relationships between performance and process variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Exog (pooled)	3.653***	3.747***	4.052***	4.170***	3.174**	2.379*	4.725***
	(1.323)	(1.356)	(1.362)	(1.333)	(1.294)	(1.267)	(1.518)
# Ideas		0.434					
		(0.665)					
# Stackings			-0.143				
			(0.206)				
# Collapses				-0.477*			
				(0.255)			
Time-to-first idea (min)					-0.542***		
					(0.161)		
Time-to-first stacking (min)						-0.465***	
						(0.120)	
Time-to-first collapse (min)							-0.275**
							(0.127)
Constant	-4.594	-5.632	-4.003	-4.031	-2.542	-5.772	-4.341
	(3.971)	(4.101)	(4.147)	(3.914)	(3.809)	(4.143)	(4.239)
Observations	112	108	108	108	108	107	90
		Variation explained by process variable					
		NA	NA	NA	17.78%	35.78%	11.23%

Note. Tobit coefficients are reported. The omitted category is the Endogenous treatment. Performance ($) is the dependent variable. Age, Engineering major (Yes/No) and gender are controlled for. The number of observations is reduced by four in columns 2-7 due to four videos being defective. Time variables are measured in minutes elapsed from the beginning of the design task. In columns 6 and 7 the number of observations is reduced due to one participant never attempting a stacking and eighteen participants never experiencing a collapse. Comparisons for which the treatment effect and the fitted value difference had opposite signs, or in which the process effect was not significant are denoted by "NA". * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Columns (2)-(4) of table 1.3 confirm that the count variables (# ideas, # stackings, # collapses) do not explain the advantage of exogenous transitions. Among the time variables (columns 5-7) time-to-first stacking has the greatest explanatory power: one minute of delay costs the designer $0.37 (marginal effect significant at $p < 0.01$) explaining approximately 36% of the performance variation. Time-to-first idea is also significantly related to performance explaining approximately 18% of performance variation ($p < 0.01$). The performance effect of first failure is somewhat weaker, explaining only 11% of the combined performance variation ($p = 0.033$). Note also

quadratic specifications, all of which confirmed our results.

that adding the time-to-first idea and time-to-first stacking to the list of regressors lowers the magnitude and the significance level of the Exogenous treatment coefficient (highest $p-$value is 0.058 in column 6). That is, the treatment effect on performance is partially explained by the delays in those activities.

In sum, early build, testing and even failures all have positive effects on design performance and explain up to one third of the treatment differences. With the endogenous decision control those activities are delayed by several minutes causing significantly reduced performance.

1.4.6 Design Process: Discussion

The analysis of the micro-structure of the design effort highlights some behavioral manifestations of the endogenous decision control. When endowed with full scheduling autonomy designers launch their first construction, attempt their first test and experience their first failure with a substantial delay, all of which leads to greatly reduced design performance.

The delays in physical activities explain a significant share of the performance advantage. In contrast, different Exogenous allocations of time to phases do not affect these design activities. Also, pure count measures (as opposed to timing) of design activities do not explain performance differences. Taken together these results confirm some conventional wisdoms and contradict another, commonly associated with good design practice. In particular, designers are often advised to "get physical fast" (go rapidly to first build) and "fail fast" (test early), both of which are supported by our results. Also, designers are often told that in the ideation stage "Quantity is Quality" (the more ideas, the better). This is not supported by our data.

We also find that the negative effects of delays in physical development do not disappear fully once the transition time is fixed exogenously. In fact the delays explain about one third of the performance difference between the (pooled) Exogenous and the

Endogenous treatments. This has two important implications for managing product development. First, managers may be able to improve new product performance by imposing strict time bounds on ideation and by limiting extensions of the exploratory period. Second, design performance may benefit from an additional requirement on design teams to build early physical prototypes of their ideas. We present new evidence for the validity of the latter recommendation in section 6 where we reexamine the effect of early physical build/test and the effects of prototyping on performance.

The benefits of early prototyping and testing have been studied extensively in the product development and project management literature (*Iansiti*, 1995; *Thomke*, 1998; *Dow et al.*, 2009; *Parvan et al.*, 2015). However, the literature is less explicit about the relationship between early physical representation of design ideas and scheduling autonomy. In fact, the literature sometimes considers "flexible development processes", "failing fast" and "get physical fast" to be part of the "lean" development paradigm and does not differentiate between decision control and process-related recommendations to design teams (c.f. *Iansiti*, 1995; *Maccormack et al.*, 2001; *Biazzo*, 2009). Our findings suggest that firms need to be cautious when applying "lean" ideas to their design projects. In our data scheduling flexibility had negative design consequences, whereas strict bounds on ideation resulted in earlier physical build and testing leading to greatly improved performance.

The above analysis and discussion leave several process-related questions unanswered. First, we cannot determine with the extant data whether the delays in time-to-first build and test are the immediate cause of poor performance (in which case they may be directly addressed by managers) or if they are a manifestation of some latent cognitive effect of the endogenous decision control. Second, while we were able to uncover several drivers of design performance the gap between Exogenous and Endogenous transition groups remains even after controlling for design activities, endogenously determined transition times and individual differences (to the extent that

we could identify and measure those factors). This suggests a more careful examination of the psychological drivers of performance differences.

Two psychological effects of endogenous decision control are suggested by the literature. Endogenous decision-making could result in the experience of choice overload (*Iyengar and Lepper*, 2000), or (more generally) cognitive overload caused by the complexity of the task (*Dennis et al.*, 1996, 1999). In particular, being preoccupied with scheduling tasks may detract from direct value-adding activities leading to poor performance. Or, alternatively a fixed transition point may provide the designer with a motivational boost by signaling the approaching phase transition. Recent work in goal-setting theory has shown that milestone progress checks that give individuals feedback on their advancement to a superior goal (here: design success) may lead to better work outcomes (*Locke and Latham*, 2002; *Fishbach et al.*, 2006). Indeed, exogenously imposed transitions may be perceived by designers as process goals improving self-efficacy and design performance.

To explore these possibilities the next section will examine three new treatments. If early building and testing is the key to better performance, we may be able to enhance performance by sharing this wisdom with designers. This could be expected to encourage earlier building and testing, and potentially enhance performance. If cognitive load is the reason for deteriorated performance, then relieving the designer of the scheduling duties by asking her to pre-commit to a transition time before the task begins should exhibit enhanced performance. If framing the process as proceeding in phases is the key to better performance, even if timing is chosen endogenously, then we should be able to enhance performance by demanding a minimally performing deliverable that clearly punctuates a phase, prior to allowing a transition.

1.5 Additional Treatments: Alternative Scenarios with Endogenous Transition

We examine three new scenarios that allow a clean test of some of the recommendations developed in section 4, help determine the relative importance of the psychological drivers of performance, and explore whether improved performance can be achieved without imposing an exogenous time schedule. In particular, we examine transition regimes in which (1) transitions are endogenous but early build and testing are encouraged, (2) transitions are endogenous but designers are asked to choose binding transition times before they start working, and (3) transitions are endogenous but are permitted only after a demonstration of a minimum performance prototype. Scenario 1 re-examines our process-related recommendation discussed in section 4 (of encouraging early build and testing).[18] Scenario 2 tests whether relieving designers of scheduling duties while they are working on the design task is the main driving force. Scenario 3 reflects a compartmentalized "stage-gate"-like regime with the design task clearly framed as a phased process.

1.5.1 Experimental design

The basic setup of the new treatments was similar to the original four treatments: subjects worked on the same design task, were given 20 minutes for completion and the task was divided into the ideation and execution phases. 95 subjects were recruited for these treatments. The treatments resembled the three scenarios described above (the instruction text is reproduced in the electronic companion, at the end of this document). In the first new treatment (henceforth referred to as the Nudge treatment) we examined the effects of encouraging early build and testing. Designers transitioned endogenously and were free to pursue any design strategy and choose the

[18]Note that sharing information with the designers may encourage a sense of urgency about certain tasks, but would not clearly frame the creative process as a phased process.

transition time as they saw fit. However, they were advised to begin early with physical build and testing. They were also informed that previous experiments indicated a positive relationship between early build/testing and performance. In the second (Pre-commit) treatment we asked designers to commit to a transition time before they began working. The third (Prototype) treatment was identical to Endog with the exception that transition into execution was allowed only after designers were able to demonstrate a minimum viable construction (worth at least \$1, corresponding to the 25^{th} percentile of the performance distribution in the original four treatments). Designers who were not able to demonstrate a minimum viable construction were not allowed to transition into execution receiving a payoff of \$0.[19]

1.5.2 Experimental results

Performance comparisons

Design performance in each new treatment is not significantly different from the (pooled) Exogenous transitions (Rank sum tests, all $p > 0.40$). Mean performance in the Nudge (Pre-commit, Prototype) treatment is \$5.53 (\$6.07, \$6.67). That is, each of the treatments is associated with improved design performance, relative to the Endogenous treatment (\$3.39). However, while requiring a prototype and asking to pre-commit to a transition time both lead to significant improvements (Rank Sum tests, $p = 0.023$ and $p = 0.025$, respectively) the advantage of the Nudge treatment is only marginally significant ($p = 0.089$).

The performance advantage of the new treatments relative to Endog is partly driven by fewer design failures. However the differences in the proportion of failures are not statistically significant (Probit regressions, $p > 0.26$). That is, Exogenous 5/15 and 10/10 are the only regimes with the failure rate being significantly reduced,

[19]In all treatments participants were paid based solely on their final performance; prototype performance was not incentivized.

relative to the Endogenous base case (c.f. columns 1 and 2 of table 1.2). Performance results conditional on non-failure are similar to the unconditional performance results. In particular, Nudge improves performance given non-failure, but not significantly (Rank Sum test, \$6.86, $p = 0.249$), whereas Pre-commit and Prototype are significantly better, relative to the Endogenous treatment (\$7.77 and \$8.90, $p = 0.021$ and $p < 0.01$, respectively)

Column 1 of Table 1.4 reports Tobit regression coefficients with performance as the dependent variable (baseline treatment: Endog). The average performance advantages of the Nudge and the Pre-commit treatments are \$2.05 ($p = 0.082$) and \$2.03 ($p = 0.078$), respectively.[20] However, the highest performance level is exhibited in the Prototype treatment (average marginal effect: \$3.20, $p = 0.010$). In columns (2)-(4) we control for the effects of the process variables that have previously been shown to affect design performance. Consistent with our previous findings one minute of delay in the first physical build (first stacking, first failure) is associated with performance drops of \$0.39 (\$0.33, \$0.31, all $p < 0.01$). After controlling for the time-to-first build, time-to-first test and time-to-first collapse Prototype retains its position as the best performing treatment with the treatment effects being significant at $p < 0.01$ in each specification. In contrast, the performance effects of Nudge and Pre-commit are less robust to inclusion of the process variables. The implications of this result will be discussed below.

Design process

The new treatments exhibit some differences in the activities engaged by the designers.[21] In particular, the number of ideas in Nudge and Pre-commit is related

[20]Pre-commit had a higher percentage of engineers (whose performance was significantly better relative to non-engineers, regardless of the treatment), which explains the discrepancy between the effect sizes and the significance levels in Rank Sum tests and those obtained in Tobit regressions. In the latter college major was controlled for.

[21]When coding the video data from the additional treatments we only recorded a subset of the original coding variables. The subset was selected based on the variables that were found to drive

to performance ($\rho = 0.321,\ p = 0.084$ and $\rho = 0.500,\ p < 0.01$). We did not see a positive relationship between idea quantity and performance in any of the remaining treatments (Prototype, Endog, Exog). That is, "Quantity=Quality" is not uniformly supported, but rather depends on the transition regime in question.

There were also some differences in the timing of the activities. The time-to-first idea is reduced by only 11 seconds in Nudge, relative to the Endogenous base case scenario (Rank Sum test, $p = 0.676$), while the time-to-first stacking is reduced by 2.35 minutes ($p = 0.081$). In contrast, neither the times-to-first idea nor the times-to-first stacking are significantly different in the Pre-commit and Prototype treatments, relative to the Endogenous base case. That is, front-loaded ideation (in the form of earlier tests) is both a unique feature of the Exogenous regime and a behavior that can be encouraged by communicating its advantages to designers.

We have seen previously that approximately one third of the performance gap between the Exogenous and the Endogenous treatments could be traced back to the process delays. We repeat the process analysis for the new set of treatments. The bottom panel of table 1.4 reports the results with the Endogenous treatment used as the comparison benchmark in each case. As before, these comparisons are based on the average delays in each treatment and the average marginal effects computed using the Tobit estimates. Comparisons for which the treatment effects and the fitted value differences have opposite signs are denoted by "NA".

We first replicate the comparison of Endog and Exog using the new estimates of the process variable effects.[22] We find the portion of the performance gap explained by the delays to be consistent with our previous results. The time-to-first stacking has

performance differences in the original four treatments (Time-to-first build, stacking, collapse, as well as the structural characteristics of the ideas). Due to the simplified coding procedure we expected less variability in the coding, so we reduced the number of coders from three to one.

[22]The percentages for Exog/Endog comparisons in table 4 are slightly different than those computed in table 3. This is driven by the differences in the marginal effects of the process variables that are estimated using the original four treatments in table 3 and the full data set (original + additional treatments) in table 4.

the strongest explanatory power accounting for approximately 1/3 of the performance differences between Endog and Exog. Similarly, approximately 1/3 of the performance advantage of Nudge over Endog is explained by the time-to-first stacking. The Nudge dummy variable becomes non-significant after the timing variable is added to the list of regressors. That is, the treatment effect of Nudge is substantially weakened after controlling for time-to-first stacking. In contrast to the Nudge treatment, the advantage of the Pre-commit and the Prototype treatments appears to be largely driven by other factors than the process delays. For both Pre-commit and Prototype the delays accounted for only about 1/6 of the performance differences.

In sum, the additional treatments confirm that early build and particularly early testing are associated with enhanced performance. However, encouraging early build and testing does not close the entire performance gap between endogenous and exogenous transitions. Similarly, reducing the cognitive load by asking designers to make an ex-ante time allocation improves performance but does not explain the entire gap. In contrast, requiring a minimum performance prototype closes the entire gap.

1.5.3 Discussion

The new treatments help refine our understanding of the drivers of the negative performance effects of endogenous decision control. In particular, our results indicate that the performance effects of early build and testing account for a significant share of the performance differences resulting from varying the decision control. That is, "Get physical fast" is supported, both as a direct contributor to performance but also as an observable manifestation of a more latent cognitive effect that one can influence with managerial regimes.

While accounting for a significant share of the performance gap nudging designers to front-load the first physical build and testing does not close all of the gap. That is, while some of treatment differences in performance are delay-driven, it is not the

Table 1.4: Additional treatments: Treatment comparisons and timing of activities

	(1)	(2)	(3)	(4)
Exog (pooled)	3.584**	3.119**	2.498*	4.264***
	(1.383)	(1.352)	(1.367)	(1.584)
Nudge	2.937*	3.120*	2.241	2.871
	(1.679)	(1.629)	(1.634)	(1.994)
Pre-commit	2.908*	3.112*	2.429	4.103**
	(1.642)	(1.594)	(1.593)	(1.874)
Prototype	4.367**	4.951***	3.856***	6.587***
	(1.684)	(1.643)	(1.645)	(2.019)
Time-to-first idea (min)		-0.490***		
		(0.114)		
Time-to-first stacking (min)			-0.416***	
			(0.094)	
Time-to-first collapse (min)				-0.400***
				(0.104)
Constant	-0.479	2.293	2.717	1.295
	(2.988)	(2.906)	(3.010)	(3.549)
Observations	205	199	198	154
		Variation explained by process variable		
Endog / Exog		17.20%	32.61%	17.36%
Endog / Nudge		3.02%	32.45%	28.92%
Endog / Pre-commit		NA	17.82%	7.25%
Endog / Prototype		NA	17.04%	NA

Note. Tobit coefficients are reported. The omitted category is the Endogenous treatment. Performance ($) is the dependent variable. Age, Engineering major (Yes/No) and gender are controlled for. Time variables are measured in minutes elapsed from the beginning of the design task. Comparisons where treatment effects and fitted value differences had opposite signs are denoted by "NA". In column 2 the number of observations is reduced by six due to four defective videos and due to two participants not being able to develop any ideas. In columns 3 and 4 the number of observations is further reduced due to some participants never attempting a stacking or experiencing a collapse.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

single dominant factor. Similarly, cognitive load remains a viable influence on the creative process, but not in isolation. Choosing ex ante and pre-committing to a time split closes some of the performance gap between exogenous and endogenous transitions, but an unexplained portion remains. In contrast, requiring a minimum viable prototype fully closes the performance gap suggesting that endogenous transitions can indeed result in good performance when the design task is explicitly framed as a phased process with a concrete deliverable punctuating the transition.

Similar to exogenous transitions, the intermediate objective to build a prototype may be perceived by designers as a process goal improving their self-efficacy and their design performance (c.f. *Locke and Latham*, 2002). The advantage of the prototype requirement may be caused by strong motivational effects provided not only by a specific goal, but also by the immediate evaluation of and feedback on the design progress.

While these alternative endogenous regimes improved mean performance each of them was associated with increased risk, relative to the exogenous treatments. In fact design failures were significantly more frequent in each treatment with endogenous transition, relative to the exogenous regimes with short and halfway transitions. That is, while risk-neutral decision-makers may choose endogenous transitions and allow transition after a demonstration of a prototype, risk-averse decision makers should avoid any regimes with endogenous transition.

Taken together our results so far suggest that the clear compartmentalization of the design process into exploratory and execution phases leads to good design performance. The phasing can be imposed either explicitly by setting the length of the phases or by demanding a prototype that exceeds a minimum performance hurdle. Our process results indicate that the quantity of ideas matters less than the timing when ideas are launched. We next investigate the role of idea quality for design performance and its contribution to design success (or failure), relative to the role of

selecting and implementing the chosen idea.

1.6 The role of idea generation, selection and implementation

We use the structural properties of ideas to group similar ideas across designers, construct a measure of idea performance and decompose individual performance into 3 components: (1) the average quality of generated ideas, (2) the ability to select the best idea, and (3) the ability to create the best representation of that idea. Each of these steps undoubtedly contributes to design performance, but it is not clear which steps are most sensitive to active management of the creative process.

The investigation of these design activities is partly motivated by the lack of experimental and empirical research on later, more physical stages of product development. The experimental results presented in this section reveal that the relative importance of ideation and execution components in fact depends on the chosen transition regime, suggesting that a focus on creative metrics alone may hide those interactive aspects.

1.6.1 Methodology

Having recorded the codes for each idea that designers attempted as well as the payoffs earned with each idea that was submitted we can characterize the creative micro-process of each designer. We begin by computing the idea quality score for each idea that was submitted, by averaging the payoffs obtained with that idea. We then use idea quality as an input for three metrics: idea generation, selection, implementation. The idea generation score is calculated as the average quality of all ideas a designer has attempted. The selection score is calculated as the difference between the average quality of the explored ideas and the quality of the submitted idea. The implementation score is calculated as the difference between one's own final

payoff and the average quality score of the submitted idea over all subjects.[23]

By construction, the sum of the three metrics is the final payoff Π_i obtained by participant i:

$$\Pi_i = \underbrace{\Big(\mathbb{E}[\Pi_j | j \in J_i]\Big)}_{\text{Quality of generated ideas}} + \underbrace{\Big(\mathbb{E}[\Pi_k | k \in K_i] - \mathbb{E}[\Pi_j | j \in J_i]\Big)}_{\text{Selection ability}} + \underbrace{\Big(\Pi_i - \mathbb{E}[\Pi_k | k \in K_i]\Big)}_{\text{Implementation ability}},$$

where J_i is the subset of participants who have submitted the ideas that i has considered. The expectation $\mathbb{E}[\Pi_j | j \in J_i]$ is taken over all ideas that i has explored and over all participants in J_i. K_i is the subset of participants who have submitted the same idea that i has submitted. The expectation $\mathbb{E}[\Pi_k | k \in K_i]$ is taken over all participants in K_i. Because the three performance metrics sum up to the participant's overall payoff we will be able to measure what percentage of the treatment difference in performance is caused by differences in the quality of generated ideas, by the difference in selection ability and/or by the difference in implementation ability.

1.6.2 Results

Our idea pool consists of 79 submitted construction ideas (counting ideas identified by at least one coder). The most popular idea was submitted 24 times in the original treatments and 17 times in the additional treatments. One traditional measure of creativity, the novelty of an idea relative to the ideas generated by others is not rewarded in our setting. In fact, there is no significant relationship between idea "popularity", i.e. the number of participants submitting an idea, and idea quality

[23] In order to compute ideation, selection and implementation scores for each subject we construct what is sometimes referred to in the innovation literature as the "idea pool" – a collection of ideas with attributes assigned to each idea, such as the number of people that engaged that idea, the idea-specific performance distribution etc. Idea pools have been used in several theoretical and experimental studies in the innovation and product development literature (*Girotra et al.*, 2010; *Kornish and Ulrich*, 2011; *Erat*, 2012; *Erat and Krishnan*, 2012).

(coder-specific Pearson correlation coefficients, all $p > 0.1$).

Next we investigate the extent to which the ability to generate ideas, to select an idea and to produce the best version of that idea drive performance differences between treatments. We use Rank Sum tests when making comparisons that involve the full sample of subjects, as well as OLS regressions, particularly when examining subject subpopulations.

Treatment comparisons

We do not find significant differences in any of the three metrics (idea generation, selection, implementation) between the three Exogenous treatments (Rank sum test: $p > 0.147$). Further, there are no significant differences along the ideation or selection metrics between the Endogenous and (pooled) Exogenous treatments (Rank sum test: $p > 0.196$). There is, however a significant difference in implementation when comparing the (pooled) Exogenous treatment to the Endogenous treatment (difference in means: 1.282, Rank Sum test: $p = 0.039$). Similar results were obtained in OLS regressions after controlling for the demographic variables. Decomposing the performance gap between the Exogenous and Endogenous treatments reveals that ideation explains 30.59%, selection explains 17.30% and implementation explains 52.11% of the overall performance differences. That is, ideation-driven metrics play a subordinate role in explaining the advantage of Exogenous transitions, relative to the implementation metrics.

We repeat the decomposition of the overall performance gap for the treatments examined in the additional treatments. The comparison baseline is the Endogenous treatment in each case. Our comparisons indicate that Nudge is associated with a marginally significant improvement in selection (Rank Sum test, $p = 0.067$), but not in ideation ($p = 0.155$) or implementation ($p = 0.255$). Neither ideation nor selection are improved in Pre-commit, relative to Endog ($p = 0.386$ and $p = 0.273$). However,

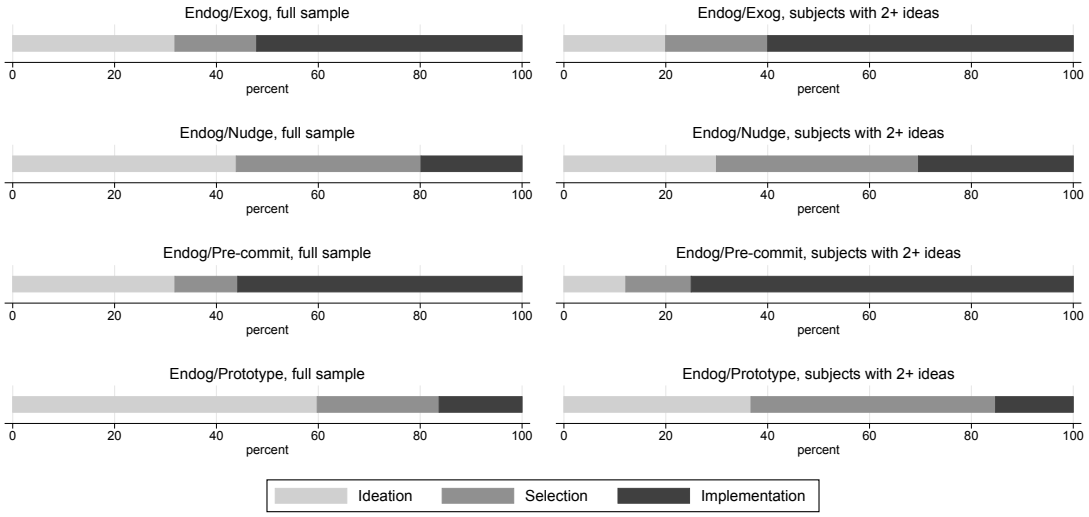
Pre-commit is associated with significantly improved implementation ($p = 0.026$). Prototype is associated with significantly improved ideation ($p = 0.020$) and improved selection ($p = 0.037$) but not implementation ($p = 0.193$). Similar results were obtained in regression analysis controlling for the individual differences.[24]

The portions of the overall performance gap explained by the three metrics are summarized in the left half of figure 1.4. To improve precision and to account for the individual differences this analysis uses OLS predicted values rather than the raw data. The contribution profiles reveal two patterns in our data. First, the portions of the performance gap explained by the ideation, selection and implementation metrics are similar in the Exogenous transitions and in Pre-commit. This suggests that the implementation advantage of Exogenous transitions is driven mainly by the ex ante allocation of the time to phases, rather than by the exogeneity of the time constraint. Second, selection and, to a greater extent ideation drive the performance advantage of Prototype with ideation explaining almost 60% of the performance gap to the Endogenous treatment. In fact, ideation performance in Prototype is significantly improved not only relative to the Endogenous treatment, but also relative to the Exogenous treatment (Wald test, $p = 0.035$).

In sum, while the treatments with ex ante fixed transition (Exogenous and Pre-commit) lead to better physical implementation of the chosen idea, treatments in which the transition decision is made "on-the-go" (Nudge and Prototype) improve the quality of ideas and the ability to select good ideas, relative to the Endogenous

[24]There were several instances when an idea was attempted but not submitted by anyone. The reported results exclude such ideas. For robustness we re-ran the analysis with an imputed score assigned to such discarded ideas. The imputation was done by regressing the mean payoffs of the submitted ideas on their structural characteristics and then by generating predicted scores for the discarded ideas. With this specification the differences between Exogenous treatments remained non-significant while the ideation and the selection advantage of Prototype remained unchanged (by construction, the implementation metrics is unaffected by the discarded ideas). We also ran the analysis considering only ideas that were submitted by at least 2 subjects to account for possible noise in unique idea quality measures. The results were similar to the reported analysis. The implementation advantage of Exog (difference in means: 1.70, Rank Sum test: $p = 0.026$) and the ideation advantage of Prototype could be confirmed (difference in means: 1.44, Rank Sum test: $p = 0.048$).

Figure 1.4: Idea generation, selection and implementation contribution to performance gap



Note. The bars indicate the shares of the performance gap explained by each of the three metrics (ideation, selection, implementation). The percentages are obtained by first computing OLS marginal effects for each metrics with Endog as the baseline and then by dividing the marginal effect on each of the metrics by the sum of those marginal effects. Age, gender and engineering major are controlled for.

base case.

Alternative metrics

To understand the role of idea selection the above analysis was repeated for the subset of designers who explored at least 2 distinct design ideas (our previous analysis may downplay the role of selection because the selection score is 0 whenever a designer explores only one design idea).

We find that restricting the sample to subjects with at least two ideas puts a greater weight on selection. Selection is driving a substantial portion of the performance gap between Endog and Nudge and of the gap between Endog and Prototype (37.93% and 43.84%; Rank Sum tests: $p = 0.155$ and $p < 0.01$, respectively). By contrast, selection explains no more than 20% of the performance advantage of Exog and Pre-commit. The right panel of figure 1.4 repeats the decomposition of the

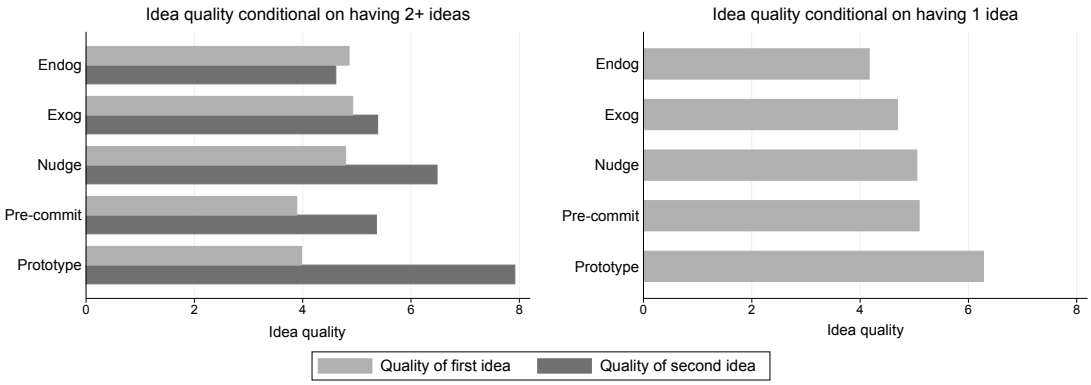
performance differences using OLS predicted values rather than the raw data. The results indicate that after controlling for the demographic variables selection explains approximately 40% of the performance gap between Endog and Nudge and 48% of the performance gap between Endog and Prototype. By contrast, selection fails to explain the performance advantage of Exog or the advantage of Pre-commit even in this restricted sample, accounting for at most 20% of the gap.

Selection (as defined in this paper) can only be a performance driver when there are quality differences between the explored ideas and the submitted idea. Therefore, it may be informative to examine the temporal sequence of ideas and the quality of each idea by treatment.

The left panel of figure 1.5 shows that Prototype exhibits a substantial quality improvement as one goes from the first to the second idea (mean difference: \$3.85, two sample $t-$test, $p < 0.01$). Subjects do not substantially improve idea quality as they explore new ideas in any treatment other than Prototype. In fact, in the Endogenous treatment subsequent ideas are on average \$0.19 worse than initial ideas. Further, designers in Prototype produce significantly better second ideas, relative to the Endogenous treatment (OLS treatment coefficient: 3.36, $p = 0.023$). However, as shown in the right panel of figure 1.5 even those designers in Prototype who explore only one idea achieve higher ideation scores relative to the Endogenous base case (OLS treatment coefficient: 2.57, $p = 0.028$). In fact, Prototype is also better than the Exogenous treatment (Wald test, $p = 0.054$).

Lastly, the quality of *submitted* ideas is significantly improved in Prototype relative to the Endogenous (OLS treatment coefficient: 3.21, $p < 0.01$) and also relative to the Exogenous treatment (Wald test, $p = 0.018$). In sum, the positive ideation effect of Prototype extends to comparisons of the initial idea, the subsequent ideas, and the submitted idea.

Figure 1.5: Idea quality by treatment



1.6.3 Discussion

Using the pool of all design ideas attempted and/or submitted in our experiment we have shown that physical execution of an idea explained most of the performance advantage of exogenous decision control. In contrast, the average quality of ideas and the ability to select good ideas was not affected by the decision control. This result is new in the literature on creativity and innovation that has been almost exclusively focusing on early ideation stages of the process. In contrast, the advantage of Prototype was driven mainly by idea quality. Initial ideas, average explored ideas and submitted ideas were all improved, relative to endogenous transitions, in fact idea quality was improved even relative to exogenous transitions. Taken together, these results suggest that the relative contribution of ideation, selection and implementation components interacts with the chosen transition regime. A focus on ideation-driven or implementation-driven metrics alone may therefore lead to poor design outcomes.

In the presence of the prototype requirement designers who attempted only one idea exhibited superior idea quality scores, relative to the Endogenous base case scenario. At the same time, those designers who did not submit their initial idea (which was typically used as the prototype) frequently had low quality initial ideas, but were able to significantly improve idea quality later on. This suggests that the prototyping

requirement may trigger a more conscious evaluation of the design approach, leading to improved allocation of the development time.

The finding that individuals are able to discover better ideas when required to prototype is also in line with the findings in the problem-solving and brainstorming literatures. Cognitive evaluation theory (*Deci and Ryan*, 1985) posits that individuals feel more competent and capable of completing a task when they experience a feeling of being "on track". Intermediate milestones may increase workers' intrinsic motivation by enabling that experience. Positive effects of an expected evaluation on creative performance have been found in a verbal task (*Shalley and Perry-Smith*, 2001). In our experiment the prototype requirement may be seen by designers as a milestone check providing progress feedback and giving them a feeling of being in control leading to improved idea quality.

1.7 Concluding remarks

This is the first experimental attempt to our knowledge to study the design performance effect of time allocations to ideation and execution phases in an innovation task, and the decision rights for choosing transition times between them. We used a controlled laboratory experiment with individual designers working on an open-ended design challenge to create a physical product subject to clear and measurable performance objectives.

The main insight from our analysis is that design performance suffers when all decisions are left in designers' hands. Imposing constraints on the design process, either in form of exogenous transition times or in form of a concrete transition point deliverable, outperformed giving designers full decision-making autonomy. This is surprising given that putting decision rights where information is richest, and giving individuals control over their work are expected to be beneficial based on the job design literature.

Another surprise was that within the Exogenous treatments we saw no significant mean performance differences in transition time. One might intuitively expect that since both ideation and execution are important, a transition at the halfway point might be best. Instead, the average performance was constant regardless of transition time, but there was a risk-return trade-off. Variance goes up with the length of the ideation period, mostly driven by a high incidence of failures for late transitions. That is, although a risk neutral firm would be indifferent, a risk-averse firm would prefer shorter ideation and longer implementation periods with the converse being true for a risk-seeking firm.

We analyzed the gap between the Exogenous and Endogenous treatments by looking at the micro-structure of the creative process, and found that the quantity of explored ideas did not consistently predict performance. The conventional logic, "Quantity = Quality when brainstorming" featured mixed results in our experiments, and is probably not uniformly true. In contrast, the timing of activities differed between the Exogenous and Endogenous treatments, at least partially explaining the results. Specifically, delays in important activities such as the appearance of the first idea, the first test and even the first failure were significantly related to poor design performance (but did not explain all of the performance differences). So "Get Physical Fast" and "Fail Fast" are robustly good recommendations, but do not in isolation explain performance gaps.

The results around rapid build/test align with conventional design wisdom, but the independent effect of an exogenous deadline is less intuitive. To better understand the advantage of exogenous deadlines we examined several alternative scenarios in which transitions were designer-determined, but the transition process or the information provided was changed. We found that delays in physical construction could be prevented by encouraging early build/testing, but that alone was not sufficient for good performance. In contrast, allowing transition only after designers were able

to present a minimum performance prototype led to performance levels on par with exogenous transitions. However, the prototyping requirement lead to significantly increased failure risk, relative to Exogenous regimes with early and halfway transition.

Given the emphasis on idea generation in the creativity literature, we attempted to separate out the impact of the quality of the ideas generated on performance, relative to idea selection and implementation. The relative contribution of ideation, selection and implementation varied by treatment, with each of them being significant in one or more treatments. So, all three can be vehicles for success or failure (and none can be ignored).

Our paper addresses the class of projects with a hard launch date, increased costs of exploring new ideas in execution, objective, easily measurable performance metrics and individual designers or strong team leaders. Our boundary contains many physical, engineering products in such industries as automotive, aerospace, medical devices, computers, industrial equipment, and component engineering for B2B products. Our findings do not directly inform other contexts, however survey data from 14 cross-functional teams (76 students with engineering, business, and art and design background) who spent 12 weeks designing and developing physical consumer products suggest that some of our findings may carry over to broader settings. The transition from ideation to execution was endogenously determined by those teams. Consistent with our findings, the two most frequently named obstacles to design success were delays in physical build (mentioned by 55% of respondents) and planning/scheduling difficulties (mentioned by 18% of respondents).[25]

Our results have several managerial implications. Managers should not endow design teams with full decision control, but rather exogenously impose a constraint that clearly signals a punctuation point between the ideation and execution phases of a creative project. Two ways to impose such external requirements are to exogenously

[25]The data and the detailed description of those design projects are presented in the electronic companion, at the end of this document.

fix transition times or to demand a concrete, performance-oriented deliverable prior to allowing the team to transition. The latter alternative is particularly relevant for product development settings in which managers are not able to set or enforce strict time schedules, or for settings where external reviews exist but transitions are de facto endogenous.[26]

Risk-averse firms will prefer exogenous transitions with longer execution times, while risk seeking firms can either impose shorter ideation times, or they can leave the decision control to design teams and request minimum performance prototypes. Regardless of the transition regime managers should both encourage and look for early build and test, because these can directly help performance as well as being markers of a productive inner design logic.

[26]There is frequently a high level of information asymmetry between a design team and the reviewers in a phase review, who are often more senior managers responsible for managing a portfolio of many projects. In such cases the potential exists for a team to strongly influence the reviewers' decisions by strategically choosing the information it presents.

CHAPTER II

Designing Incentives in Startup Teams

2.1 Introduction

In the presence of tightly constrained cash flows startup founders are frequently compensated through equity shares—the rights to participate in the proceeds from going public or from being acquired by another company. The division of equity among the founders is manifest in their first "term sheet" that specifies how and when equity vests (becomes earned) and under what conditions it can be withheld. The design of these provisions has attracted considerable attention in the entrepreneurial press with the conventional wisdom suggesting that equal splits are poor choices (*Wasserman*, 2012; *Moyer*, 2012). The conventional logic is that by not connecting rewards to either effort or contribution level equal split contracts can encourage free-riding behaviors. In this paper we experimentally test this conventional wisdom, among other entrepreneurial contracting hypotheses. In particular, we explore two research questions: what is the effect of (1) contract form and (2) contracting time on founder effort and on the value generated by the startup team?

Equity division in startups can take many forms from equal division to contribution-proportional splits. The latter are based on the value assigned to various inputs provided by the founders including labor, capital and other assets, as well as contacts and business leads they bring to the team (*Moyer*, 2012). Incentive theory suggests

that such arrangements can align individual self-interest with the firm's objectives better than equal division rules, predicting that they will result in higher effort levels and value generation.

While the incentive strength arguments favor non-equal contracts, some of their theoretical benefits may not be realized in practice, or at least not to the fullest extent. Indeed, it is often unclear what team members generate value, and how much. Many technology startups log working hours and track the number of lines of code written, however both measures are crude gauges of effort and poor predictors of value generation (*Graham*, 2004). Also, the significance of some key events may be apparent ex post but may not be easily recognizable at the time those events occur. Examples are industry contacts that open up new markets or product features whose functional appeal is not apparent until a complementary technology emerges. The delays in realization of an input's true value, and the interactions between various inputs make it difficult to evaluate each contribution separately reducing the appeal of contribution-based contracting.

Indeed, in practice most startups avoid including detailed effort or value tracking into their term sheets. Instead, many prefer simpler contracts that include minimal performance-oriented contribution thresholds, frequently referred to as "vesting" contracts (*Metrick and Yasuda*, 2010). In these the initial equity allocation is tentative and the final splits are granted only after the team members satisfy some pre-specified contribution requirements. When a vesting requirement is not met, the unvested shares are withheld and redistributed to the remaining shareholders. By ignoring minor differences in contribution amounts vesting contracts therefore serve as a compromise solution between equal and proportional contracts.[1]

[1] A common form of vesting provisions is time-based vesting, in which team members earn shares by simply remaining part of the team for a specified length of time. An alternative approach to time-based vesting is milestone-based vesting, in which the split is confirmed only after some milestone tied to individual contribution has been reached. A milestone is typically an event that correlates with value creation, such as the completion of a prototype or the first customer shipment. For sample vesting contracts used in practice see *Metrick and Yasuda* (2010).

Simpler, more egalitarian contracts are supported not only by practical considerations (of not being able to track contributions), but also by the evidence of their motivational benefits. Specifically, the human resources and the behavioral economics literatures both suggest that large differences in earnings within the team may lead to undesirable behaviors. Sharing the risks and the rewards equally may emphasize solidarity with collective interests and promote cooperative behavior (*Morgan and Sawyer*, 1979; *Deutsch*, 1975; *Kroll et al.*, 2007), while large differences in pay may lead to adverse reactions depressing effort and contributions (*Pfeffer and Langton*, 1993; *Fehr et al.*, 2009). If these arguments apply in the startup contracting context, contribution-proportional equity division should be avoided and vesting contracts may be preferred, to allow some redistribution of equity while guarding against excessive free-riding.

While there are few empirical studies on startup contracting, some survey data suggest that equal division is associated with lower outside investments and with lower valuations, relative to non-equal splits (*Hellmann and Wasserman*, 2016). However, there are some important differences in the characteristics of teams choosing different contracts. Equal contracts are preferred by family firms whose ventures are frequently funded by their informal networks and not by outside investors (*Sahbaz*, 2013; *Hellmann and Wasserman*, 2016). Contractual choices are also affected by founder experience with more seasoned teams including more contribution-dependent components into their contracts. These selection effects dominate the relationship between contract form and startup performance, leaving open whether the effect of contract form on value creation is causal.

The timing of the contractual agreement is another important consideration for incentive design in practice. Frequently the equity terms are not negotiated until part way downstream in the innovation process. In particular, equity agreements are often made at certain milestones, such as the conceptualization of the business

idea, internal or external funding events, or the start of business operation (*Jared*, 2016). When founders contract in the very early stages (i.e. before finalizing the product concept), the direction of the venture and the roles of the founders are often uncertain. In contrast, when founders contract after some work is completed, at least some of the uncertainties will be resolved prior to contracting. This may lead to better informed contracting decisions and to greater satisfaction with the contract, increasing effort and value generation (*Wasserman*, 2012).

However, delayed contracting also has its drawbacks. The human resource literature suggests that pay ambiguity may reduce worker motivation (*Belt and Paolillo*, 1982; *Barber and Roehling*, 1993; *Yuce and Highhouse*, 1998). Not knowing how their efforts will be rewarded the team may be reluctant to commit to the startup needs prior to contracting. In particular, early-stage developers may feel discouraged from participating in value creation if they anticipate that their efforts will not be fully reflected in the contract.

To understand the effects of contract form and contracting time on founder effort and on startup performance we develop a new experimental game that captures several key elements of the entrepreneurial innovation process. The value creation begins with the founders jointly determining the initial startup value by deciding how much effort to exert. Then, after observing the value generated in stage 1, the effort allocation decisions are repeated in stage 2. The individual contributions and the final value of the startup are correlated with founders' effort investments but are also affected by random noise. Once the final value is known, it is divided between the founders according to an allocation rule (as will be explained below).

Our experimental investigation allows endogenous contract selection among several contract alternatives that parallel the contract forms used in practice. These include equal, vesting and proportional division rules. To isolate the incentive effects from the effects of the negotiation process and of selection we conduct control

treatments in which we impose the contract form exogenously. After studying the effects of contract form on contribution behavior we examine whether delaying the contracting until after stage 1 affects founder efforts and startup value.

This is the first study to our knowledge to investigate the effects of equity contracting on effort and value creation in an experimental setting.[2] More broadly, this is one of the first experimental studies of incentive design for (and by) collaborative teams in the innovation and technology management literature, which often treats incentive design in technology projects as a "principal-agent" problem bypassing any within-team interactions (*Loch*, 2017).

Our results confirm the relationship between equal splits contracting and depressed effort and contribution, but suggest a different causal sequence relative to conventional wisdom. Rather than the contract form being the primitive and the behavior the derived consequence, our results suggest the reverse. Personal characteristics are the primitive and the contract form the derived consequence. In particular, our data reveal the presence of three behavioral types (low, conditional and high contributors) that differ in their preferences and behaviors. When contracting happens upfront low contributors select into equal contracts and the remaining types select into non-equal contracts. This results in the free-riding behaviors occurring more frequently in equal contracts relative to non-equal contracts. That is, equal contracts are bad for team performance, not because of their incentive strength but because of the founder types that self-select into them.

However, when contracting is delayed, teams operate with richer information when deciding on the contract. Free-riding intent of low contributors is revealed early on, and others do not want to sign equal contracts with them. Further, robust con-

[2]The only existing experimental studies on equity contracting known to the authors are *Jared* (2016) and *Bao and Wu* (2017). *Jared* (2016) explores the effects of contracting time on norm formation (cooperative vs. competitive norm) and focuses on equal splits. *Bao and Wu* (2017) examine inequality attitudes of employees to differences in equity and in salary. Our study is different because it focuses on startup teams, explores the effects of both contract form and contracting time, and because it examines effort as the main dependent variable.

tributors are also revealed early on, which reduces others‘ reluctance to sign equal contracts with them. Together, these behaviors result in low contributors no longer being over-represented in equal contracts. More generally, since it is founder type rather than the contract type (strength of incentives) that primarily impacts behaviors, with a stronger signal of type the contract form becomes less important leading to a more even distribution of types over contracts and to smaller effort and value differences between contracts.

Our findings have implications for startup investors and founders. Our results add texture to the conventional wisdom that investors should avoid startups with equal split contracts, clarifying that this result is driven primarily by the personal characteristics of the teams selecting different contracts. Both investors and founders should pay as much (or more) attention to personality type as they do to contract form. But, if one is stuck with a given set of personalities delayed contracting (more so than contract form) can improve performance.

2.2 Literature

There are several streams of literature that are relevant to our investigation. We will first discuss the empirical research on the effects of equity splits on firm performance and then move to the broader behavioral and experimental economics literature on incentive design in collective production settings.

2.2.1 Entrepreneurship literature

Given the theoretical arguments in support of input or contribution-based contracts as effective incentive instruments one may expect to find many startups using such contracts. However, the contrary is the case in practice: equal division rules are used frequently by startups and by partner-owned firms, more generally. *Encinosa et al.* (2007) find that 54% of small medial-group practices divide all profits equally.

Farrell and Scotchmer (1988) present similar data for law partnerships. *Jared* (2016) reports that 64% of South-East Asian startups have an equal ownership split between founders. *Hellmann and Wasserman* (2016) survey North American technology startups and find that 35% divide equity equally.

To our knowledge, the only study to examine empirically the relationship between equity splits and startup performance is *Hellmann and Wasserman* (2016). The survey-based evidence therein suggests that equal contracts are associated with reduced outside investment and with reduced VC involvement. However, the authors do not find a causal link between the contract form and those metrics. Rather, they argue that equal contracts are chosen by teams with close social ties who tolerate reduced team effort and value generation in favor of greater income equality. Our data confirm that a large proportion of teams reject contribution-proportional splits and that profit-seeking is not the sole motive for many teams, but suggest a different mechanism. While some individuals are indeed driven to equal splits by inequality aversion, a preference for equal contracts is most strongly associated with the desire to free-ride on partner effort.

Other empirical and experimental research also questions the incentive strength argument. *Kroll et al.* (2007) show that a more egalitarian division of shares between the founders improves startup's post IPO performance. Their argument is based on increased team cohesion in groups with an even ownership structure. The team cohesion argument is broadly related to the literature on horizontal pay differences showing that productivity may suffer as a result of unequal pay (*Pfeffer and Langton*, 1993; *Fehr et al.*, 2009).

Finally, some entrepreneurship research indicates that the focus on incentive strength of the contract may hide some interactive aspects that are relevant for startup performance in practice. *Breugst et al.* (2015) explore the collaborative dynamics in a case study of 8 entrepreneurial teams some of which have equal and some non-equal

contracts. They find that it is not the equity split per se, but its perceived justice that affects team interactions and team effectiveness. In a similar vein, *Jared* (2016) shows that equal splits may lead to a conflicted or to a cooperative environment depending on the contextual circumstances of the equity negotiations.

These findings give some insights into the sociological and psychological antecedents of a team adopting (or not) equal contracts, but provide little advice for startup teams. In our investigation we are able to study both the incentive effects of contracts and the selection effects, by examining scenarios with endogenously selected and exogenously imposed contracts. Further we focus on the direct effects of contracts on effort and contribution dynamics bypassing the contextual details of founder-investor negotiations that may interact with the effects of contracts on cooperative behavior.

2.2.2 Behavioral economics literature

The micro-foundations of contribution behavior in team settings have been studied in the behavioral economics literature, particularly in the context of public goods provision. One robust finding is the reduction of free-riding in regimes allowing punishment of low contributors (*Ostrom et al.*, 1992; *Rapoport and Au*, 2001; *Gürerk et al.*, 2006, 2009; *Gürerk*, 2013; *Putterman et al.*, 2011). *Engel* (2014) examines mild and harsh punishments and finds that the positive effect of punishment on contributions increases in the severity of the punishment. If these results carry over to the startup setting, we should see proportional contracts perform best and equal contracts perform worst. However, one caveat to extrapolating these findings to our setting is the reward allocation system used in the public goods studies. These typically assume voluntary punishment by group members, whereas startup teams use contractual sanctions.

Several studies suggest that when effort decisions are private or when effort cannot be observed perfectly, the advantage of high-powered incentives may collapse (*Cappe-*

len et al., 2007, 2010; *Fischbacher*, 2007; *Grechenig et al.*, 2010; *Bornstein and Weisel*, 2010; *Sousa*, 2010; *Ambrus and Greiner*, 2012). These papers show that teams are willing to punish low contributors only when the differences in contribution amounts are caused by free-riding and not when caused by luck. Indeed, *Bao and Wu* (2017) show that workers are more sensitive to arbitrary differences in equity compensation, than in salary compensation. Further, while both profit maximization and equitability are important concerns, a significant share of individuals split equally in order to signal unity to their partners (*Corgnet et al.*, 2011; *Luhan et al.*, 2013). If both profit seeking and fairness concerns are important determinants of behavior in our setting, vesting contracts may outperform both equal and contribution-proportional division rules.

A related set of studies examines whether individuals who exhibit socially desirable behaviors select into less egalitarian reward allocation regimes. *Balafoutas et al.* (2013) find that low contributors select into regimes with redistribution, but the selection effect is dominated by incentive effects. *Tyran and Feld* (2006); *Gürerk et al.* (2009) and *Sutter et al.* (2010) show that selection effects can be stronger than incentive effects in the public goods game setting. In a prisoner's dilemma game with and without punishment *Dal Bó et al.* (2010) show that both incentives and selection affect the frequency of defections. Consistent with these results we find that the preference for equal splits is associated with free-riding behaviors, and that the sorting of low contributors into equal contracts is the primary driver of contract performance differences. However, we also find that the extent to which free-riders are able to select into egalitarian regimes depends on the availability (or lack) of effort information prior to contracting.

The existing empirical and experimental literature is relatively silent on the effects of contracting time on startup performance. *Sahbaz* (2013) and *Hellmann and Wasserman* (2016) report that a non-trivial share of startups delay contracting until

further downstream in the innovation process. However, they do not find a significant relationship between contracting time and performance. *Wasserman* (2012) argues that early contracting may create clarity around the incentive structure, increasing effort levels. Early negotiations may lead to fewer conflicts among the founders, particularly if the stakes increase over time. However, *Wasserman* (2012) also notes that delaying the contracting may reduce the uncertainty around the firm value and the individual contributions to it. This may help craft a more informed and thus a more effective contractual agreement. *Jared* (2016) finds that delayed equal splits lead to more cooperative norms relative to upfront equal splits. Though *Jared* (2016) does not examine the effects of contracts on effort, his findings anticipate one of our results, that contract performance depends on the availability of mutual effort signals prior to contracting.

In sum, the extant empirical research presents mainly correlational evidence and mixed results. The experimental literature suggests that allowing teams to penalize free-riders will lead to higher contributions and value creation. This supports contribution-based contracting. However, by focusing on one-shot contribution decisions, observable efforts and (predominantly) ex post division of the surplus these experimental studies are only partially reflective of the entrepreneurial context. None of the existing experimental studies provide clear recommendations for entrepreneurs, partly because the division rules examined there do not resemble the contractual agreements used by startups in practice. Our model and experiment are designed to address this gap by following more closely the contracting and collaborative dynamics in startups.

2.3 A stylized model of entrepreneurial contracting and value creation

The contracting and collaboration environment in our experiment reflects several features shared by many entrepreneurial ventures. The following scenario is the stylized context of our model that maintains the relevant features and is used in the experiment. After introducing the model setup we will discuss equilibrium effort levels implied by each contract form.

2.3.1 Setup

A startup team consisting of two partners has identified a problem that they want to develop a product to solve, creating a new business that they will own. They do not yet know what the actual value of the business will be, or how much effort each partner will allocate to the venture. There are two phases to the business development effort. In each phase, each partner can choose to invest effort in the venture (with a risky return as described below) or an outside option (with a certain return). This is to model the outside employment or other options that each individual has, which is also the opportunity cost for the effort invested in the venture. In practice, phase I may feature market research, product concept selection and product development activities, while phase II may involve more downstream processes, such as setting up the supply chain or marketing and sales activities.

Each partner $i \in \{1, 2\}$ begins stage $s \in \{1, 2\}$ with a finite effort endowment E that she can allocate between the venture and the outside option. There are two dimensions to the real value increase of the venture as a result of the cooperative efforts of the founders. First, each founder chooses to contribute effort $e_{is} \in [0, E]$ to the venture. Second, the venture value is increased based on the joint investment of both partners. This latter mapping is uncertain. For example, effort can be expended

at a high cost to the contributor, but with a low value for the venture. However, the real venture value increment is positively correlated with the joint effort investment of both partners. Formally, i's contribution in stage s, $c_{is} = m_{is} \times e_{is}$, where m_{is} are i.i.d. discrete random variables that can take a low, medium or high value with some known probabilities (We choose a simple three-point mapping of effort to contribution to make the game more accessible to the experimental subjects). In contrast, the return to each individual for effort invested in their outside option is certain. That is, in each stage founder i earns an additional private payoff of $(E - e_{is}) \times K$, where K is a constant.[3]

Effort is private information, but the value contribution is public. After each phase each team member observes the value increment resulting from their own and from their partner's effort allocation decision, but not the partner's effort level. That is, the amount of effort actually invested by the partner is shared in form of a noisy signal. The quality of the product concept V_1 (determined at the end of stage I) depends on how much effort (and the returns to that effort) is invested in understanding customers and designing for their needs. At the end of stage I the team members see a business valuation number V_1 that is positively correlated with their joint contributions, and also has a positive signal value about what the final business value will be. In particular, $V_1 = c_{i1} + c_{j1}$. This is to model the end of the market research phase, where the potential market valuation of the business is known if the team can deliver a product or service that responds to the needs discovered in stage I.

The partners then (privately) choose their individual level of effort in phase II and the process repeats yielding stage II value $V_2 = c_{i2} + c_{j2}$. This is to model the incremental increase of the firm value resulting from the actual product launch and sales activities. At the end of stage II the team gets a final business valuation V

[3]Our model and experiment abstract away from any ex ante skill asymmetries within the team. That is, m_{is} has the same probability distribution for each partner i and in each stage s.

that is positively correlated with the value at the end of stage I, and with the joint contributions realized in stage II. In particular, the final valuation of the startup, $V = V_1 \times V_2$. This is to model the actual business value, after the product design and launch. The final earnings of each founder include her share of the firm value V and her private payoffs. That is, founder i's profit $\pi_i = \sigma_i^X V + K(2E - e_{i1} - e_{i2})$, where σ_i^X denotes the share of the startup value allocated to founder i under contract X (contracts will be discussed below).[4]

In sum, we model the entrepreneurial innovation process as a two-stage game. Value creation begins with the founder contributions jointly determining the initial startup value. These contributions are correlated with founders' effort investments but are also affected by random noise. Higher initial startup value increases the attractiveness of contributing to the startup in the second stage. Once the final value of the startup is known, it is divided between the founders according to an allocation rule. The allocation can be made contingent on the individual contributions (effort is not observable so cannot be contracted on) with four contract forms to choose from: Equal split, Threshold vesting, Difference vesting and Proportional contracts. The specifics of these contracts are described next.

2.3.2 Contracts

Our investigation focuses on contractual division rules in which the differences in future (and not past) contributions can be contracted on. Such symmetric, forward-looking contracts are typical for early stage ventures formed by teams of peers (rather than entrepreneur-adviser or inventor-employee teams) in which founder roles are comparable in importance. The contract menu used in our model and experiments

[4]Our two-stage model draws on the idea that the value of the venture is often much lower and much more uncertain before the startup has found its product-market fit.The value crystallizes once a working business model has been found. The two-stage model can also be interpreted as an abstraction to the milestone-driven growth typical for many startups. Indeed, a startup's valuation is often shown to increase at isolated and well-defined events, a proof of concept of the core technology, a successful demonstration of prototype performance, or a key customer acquisition (*Nachum*, 2015).

draws on the equity contracts used by startups in practice (*Wasserman*, 2012; *Moyer*, 2012). The contracts are further validated in pilot experiments, in which we allow teams to design their own contracts from scratch (The pilot is described in section 4.2, the contract transcripts are reproduced in the online supplementary documents, http://webuser.bus.umich.edu/ekagan/research.html).

The contract alternatives in our model and in our experiments are Equal split (henceforth EQUAL), threshold vesting (THRESH VESTING), difference vesting (DIFF VESTING) and contribution-proportional split (PROPORTION). With THRESH VESTING a player loses 10 percentage points of equity each time she contributes less than a fixed contribution threshold c^{THRESH} and the partner contributes at least the threshold amount. The lost portion of the equity is reallocated to the partner. With DIFF VESTING a player loses 10 percentage points each time she contributes less than her partner and the difference is at least c^{DIFF}. The lost portion of the equity is, again, reallocated to the other player. With a PROPORTION contract a player's share is computed as the ratio of the sum of her contributions to the sum of all individual contributions.

The contractual share allocated to player i under contract X is denoted by σ_i^X, where $X \in$ {EQUAL, THRESH VESTING, DIFF VESTING, PROPORTION}. The contractual share allocated to player j, $\sigma_j^X = 1 - \sigma_i^X$. In the UPFRONT contracting scenario equity shares are calculated as follows:

$$\sigma_i^{EQUAL} = 0.5$$

$$\sigma_i^{THRESH\ VESTING} = \begin{cases} 0.3 & \text{if } \{c_{is} < c^{THRESH} \wedge c_{js} \geq c^{THRESH}\} \text{ in both stages } (s = 1, 2) \\ 0.7 & \text{if } \{c_{is} \geq c^{THRESH} \wedge c_{js} < c^{THRESH}\} \text{ in both stages } (s = 1, 2) \\ 0.4 & \text{if } \{c_{is} < c^{THRESH} \wedge c_{js} \geq c^{THRESH}\} \text{ in exactly one stage} \\ 0.6 & \text{if } \{c_{is} \geq c^{THRESH} \wedge c_{js} < c^{THRESH}\} \text{ in exactly one stage} \\ 0.5 & \text{otherwise} \end{cases}$$

$$\sigma_i^{DIFF\ VESTING} = \begin{cases} 0.3 & \text{if } \{c_{js} - c_{is} \geq c^{DIFF}\} \text{ in both stages } (s = 1, 2) \\ 0.7 & \text{if } \{c_{is} - c_{js} \geq c^{DIFF}\} \text{ in both stages } (s = 1, 2) \\ 0.4 & \text{if } \{c_{js} - c_{is} \geq c^{DIFF}\} \text{ in exactly one stage} \\ 0.6 & \text{if } \{c_{is} - c_{js} \geq c^{DIFF}\} \text{ in exactly one stage} \\ 0.5 & \text{otherwise} \end{cases}$$

$$\sigma_i^{PROPORTION} = \frac{c_{i1} + c_{i2}}{c_{i1} + c_{i2} + c_{j1} + c_{j2}}$$

2.3.3 Model parameters

The parameters in our model and experiments are chosen such that (1) there is a prospect of a substantial (but risky) gain for both value generation and expected founder profits if the partners both invest full effort into the startup, and (2) different contract types exhibit different incentive and allocation properties, rendering the contracting and effort decisions consequential. These considerations led to the following parameter choices. Subjects are endowed with an effort budget $E = 10$ in each stage. The returns for effort, m_{is} can take values 0.5, 1, and 2. The realization probabilities of these values are 0.25, 0.5 and 0.25, respectively (m_{is} has the same probability distribution for each partner i and in each stage s). The constant multiplier on the private investment, $K = 5$. The vesting thresholds c^{THRESH} and c^{DIFF} are both equal to 5. These parameter choices were validated in a pilot with 50 subjects (The pilot is described in sections 4.1-4.2).

2.3.4 Equilibrium strategies

We next outline the equilibrium strategies in the UPFRONT and IMPOSED scenarios (The predictions for the DELAYED scenario are postponed until section 5). A more detailed description of the equilibrium structure is relegated to the online supplementary materials.

With EQUAL contracts stage II best response strategies depend only on the value of V_1. If $V_1 > \overline{v}_1$ investing full effort into the startup is the best response to any partner action. If $V_1 < \overline{v}_1$ investing no effort is the best response to any partner action. Following the backward induction logic and plugging in the continuation payoffs into the stage I profit function, "Invest full effort endowment" is the unique stage I best response to any partner action. The reason is that each player can unilaterally achieve that $V_1 > \overline{v}_1$ with a sufficiently high probability, making the expected returns for effort invested in the startup greater than the returns for the outside option. Further, because in equilibrium both partners will invest full effort in stage I, the "high" state with $V_1 > \overline{v}_1$ will always be reached in stage II resulting in full stage II effort. That is, any less-than-maximal effort investment in either stage implies off-equilibrium behavior.

With NON-EQUAL contracts the best response in stage II generally depends not only on the sum of stage I contributions (as was the case for EQUAL), but also on the individual stage I contributions c_{i1} and c_{j1}. However, it can be shown that simple strategies still exist for a range of V_1 values. Intuitively, because NON-EQUAL contracts tie the allocation of equity to individual effort and contribution, they lead to more socially desirable behaviors (i.e. equal or higher effort levels conditional on V_1), relative to EQUAL contracts. Indeed, plugging in the continuation payoffs and solving for the best response it can be shown that "Invest full endowment" is the unique best response in stage I in each NON-EQUAL contract. That is, given our parameter choices, effort differences among the contracts are predicted only in stage

II and only on the off-equilibrium path. In equilibrium each contract is predicted to lead to full effort investment in both stages (The off-equilibrium strategies in stage II are characterized more explicitly in the supplementary materials, see table S1.2).

2.4 Experimental setup and results

2.4.1 Experimental strategy

To investigate the effects of equity contract form and of the contracting time on effort and on startup performance we conducted a pilot treatment and 4 between-subject treatments labelled IMPOSED EQUAL, IMPOSED PROPORTION, UPFRONT and DELAYED. In the IMPOSED treatments contracts were imposed exogenously by the experimenters. In the UPFRONT and DELAYED treatments contracts were selected endogenously by the team.

In the pilot treatment (conducted ahead of the remaining treatments) we asked subjects to design their own contracts. The purpose of the pilot treatment was to explore inductively the contractual arrangements emerging from free-form negotiations, to validate the model parameters and to examine the frequencies of different contracts in a face-to-face setting. In the remaining treatments subjects interacted via the z-Tree interface (*Fischbacher*, 2007). In the Endogenous negotiations treatments subjects chose jointly one of four contract types (EQUAL, THRESH VESTING, DIFF VESTING, PROPORTION). Our contract menu draws on the contractual agreements used by startups in practice and also aligns with the contract types emerging from the free-form pilot, as will be described below.

2.4.2 Pilot: Free-form negotiations

50 subjects were recruited at the University of Michigan to participate in the pilot treatment. Two-person teams were formed at random, and each team was given

two empty sheets of paper to be used for writing down the contracts. Each team held private negotiations in a separate room with no time restriction. Once a team completed their negotiations, the experimenters verified that each partner had signed a copy of the contract and that the copies were identical, and brought the team to the laboratory where they continued with the contribution phase of the experiment. Each subject participated in three rounds of the startup game, with random re-matching in each round. The average duration of one negotiation round was 3 minutes and 40 seconds. On average, subjects spent one hour in the laboratory earning $14 including the show-up fee of $5.

Our pilot data show that equal splitting is a particularly appealing contract form with 73% of the teams choosing equal split contracts. The appeal of equal contracts is consistent with the behavioral economics literature on face-to-face interactions in joint production and bargaining settings (*Roth*, 1995; *Bochet et al.*, 2006; *Corgnet et al.*, 2011; *Konow et al.*, 2009) and is also consistent with the empirical entrepreneurship literature (*Hellmann and Wasserman*, 2016; *Breugst et al.*, 2015; *Jared*, 2016)

In addition to the popularity of equal contracts, we were able to identify several categories among the non-equal contracts emerging from the negotiations. In particular, the non-equal contracts fell into 3 categories: "threshold-based" (a contribution below X points is penalized, where X is a constant), "difference-based" (a contribution below Y points, where Y depends on partner's contribution), and "proportion-based" (each partner is allocated a share of the profit proportional to the share of points contributed). These allocation rules are consistent with the endogenously designed redistribution schemes in the public goods literature (*Rockenbach and Wolff*, 2016) and can also be mapped to the contracts used by startups in practice (*Metrick and Yasuda*, 2010).[5] The contract types were further refined and calibrated by the

[5] *Rockenbach and Wolff* (2016) report that endogenously designed allocation rules in public goods games are typically based either on either absolute or relative thresholds: "Mechanisms were [designed] in the form of pre-specified rules of deduction and/or redistribution contingent on complying with provision targets. These provision targets were either fixed levels (e.g. full provision) or con-

authors and then used to design the contract menu for the remaining treatments. The transcripts of all contracts written by the subjects in the free-form treatment are reproduced in the supplementary documents.[6]

2.4.3 UPFRONT and IMPOSED treatments

104 subjects were recruited at the University of Michigan to participate in the UPFRONT and IMPOSED treatments of the experiment. After going through the instructions all subjects were required to complete a mandatory quiz. Subjects then played eight rounds of the startup game, with random re-matching in each round. During the negotiations the interaction between the partners was limited to making, rejecting and accepting contract offers (subjects could not exchange chat messages). In the first two rounds subjects were given 4 minutes to agree on a contract. In the subsequent rounds subjects were given 2 minutes to agree on a contract. If a team was unable to agree on a contract, their endowments were allocated automatically to their private accounts. On average subjects spent 50 minutes in the laboratory and earned $14 including the $5 show-up fee. The exact transcript of the instruction text, and the screen shots of the negotiation screens are reproduced in the supplementary materials.[7]

In the remainder of section 4 we examine effort and value generated in each contract when the contract is determined upfront. Section 5 examines the delayed contracting scenario. Section 6 investigates the motives and behaviors of different personality types present in our data.

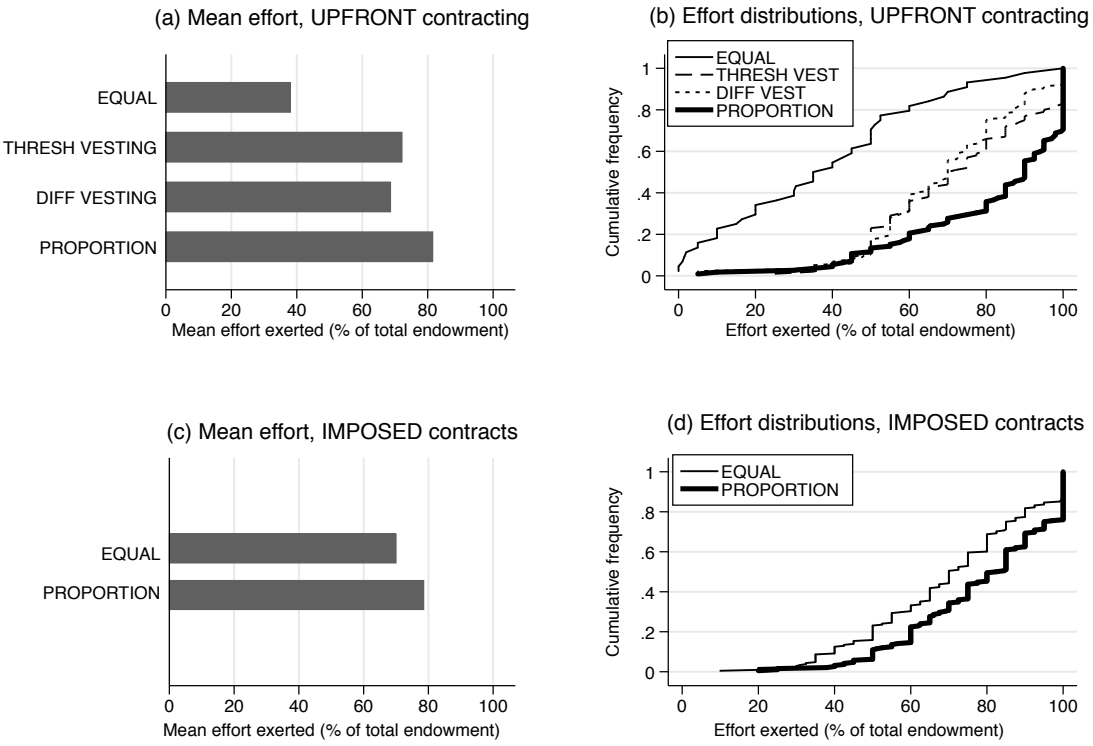
---

tingent on the other group members (e.g. not being the lowest contributing player)."(*Rockenbach and Wolff*, 2016, p. 332).

[6]Due to a small number of observations and due to a large number of different contractual arrangements we do not explore in detail the effects of contracts on effort in the pilot data. On average (pooled) non-equal contracts were associated with an increase in contribution levels of 8.60 percentage points relative to equal contracts, but the effect was not statistically significant.

[7]All but one team in our data were able to agree on a contract. Among the four contracts no single one attracted more than 40% of the teams in any given round, and each contract was chosen by a non-trivial share of teams in each round. Further, neither the negotiation time, nor the number of exchanged offers were predictive of effort and value generation.

Figure 2.1: Mean effort levels and effort distributions (UPFRONT and IMPOSED treatments)



Results: aggregate effort and value comparisons

Figure 2.1(a) shows a substantial gap in effort levels between EQUAL and each NON-EQUAL contract and a smaller gap between PROPORTION and each VESTING contract. Average (stage I + stage II) effort levels are lower in EQUAL relative to PROPORTION (mean difference: 43.49 percentage points), and also lower in pooled VESTING relative to PROPORTION (mean difference: 11.62 percentage points). Effort levels do not differ substantially between the two VESTING contracts. These results suggest that effort levels rise monotonically in the extent to which the share allocation is tied to contribution differences.

Not only the means, but also the distributions of effort levels in the UPFRONT negotiation scenario differ between contracts, as shown in figure 2.1(b). In particular,

each NON-EQUAL contract dominates EQUAL contracts in terms of effort and total value generated (V) in the sense of first order stochastic dominance (all $p < 0.01$).[8] This suggests that an investor would prefer to fund a team with a non-equal contract regardless of risk sensitivity. Further, figure 2.1(b) suggests that the advantage of PROPORTION contracts relative to VESTING contracts is driven by frequent near-maximum effort levels in the former. Indeed, over two thirds of the observations in PROPORTION contracts feature effort levels above 80% of the subject's endowment, compared to only one third of the observations in VESTING contracts.

Compared to the robust differences in contract performance with endogenously selected contracts, the differences between exogenously imposed contracts are small. In particular, figure 1(c) shows that the effort gap between PROPORTION and EQUAL contracts is approximately 9 percentage points. These results suggest that it is not the incentive structure of the contract that matters most for contract performance, but the personal characteristics of those who select these contracts.

Non-parametric tests of effort level differences

We have so far examined average effort and value generated in each contract without specifying whether multiple observations of behavior in a contract came from one subject or from multiple subjects. To isolate between-subject differences in behavior we next examine effort levels observed in a single round of the experiment.

In the *first* experimental round of the IMPOSED treatment, the effort gap between EQUAL and PROPORTION contracts is 0.20 percentage points (Rank Sum test, $p = 0.868$). In the *last* experimental round EQUAL falls behind PROPORTION by 7.34 percentage points, with the difference not being statistically significant ($p = 0.183$). In contrast, the *first* round comparison in the UPFRONT negotiation treatment reveals a 27.56 percentage point gap between EQUAL and PROPORTION ($p = 0.018$).

[8] This result is obtained using tests based on quantile regressions discussed in *Ng et al.* (2011).

Further, that effort gap widens over time reaching 59.95 percentage points in the *last* experimental round ($p < 0.01$). These results suggest that the endogenous contracting environment generates a persistent effort gap between the contracts. Further, the increase of the gap over time suggests that the differences are driven in part by observed partner behaviors, and not by the incentive strength of the contract.

Regression analysis

We next examine the effects of contracts on effort using random effects regressions. Columns 1-4 of table 2.1 report the effects of PROPORTION contracts on effort when contracts are imposed exogenously by the experimenter (baseline is IMPOSED EQUAL). The coefficients describe the changes in effort levels caused solely by the change in the incentive structure and are free of any selection effects. Column 1 shows that stage I effort levels differ by approximately 8 percentage points between EQUAL and PROPORTION contracts, with the difference being marginally significant ($p = 0.067$). Column 2 shows that this effort gap expands to approximately 10 percentage points and becomes statistically significant as we move from stage I to stage II ($p = 0.027$).

Column 3 shows that some of the effort level differences in stage II are explained by the differences in V_1. This is consistent with our equilibrium predictions. However, column 4 shows that most of this effect is driven by the responses to stage I partner contribution, c_{j1}. Column 4 breaks V_1 into some of its components and shows that one point increase in stage I partner effort is associated with 0.64 percentage point increase in own stage II effort ($p < 0.01$). In contrast, people are not sensitive to exogenous changes in the returns for investing effort measured by their stage I multiplier, m_{i1} ($p = 0.872$). These results suggest that subjects respond to incentive strength differently than suggested by standard theory, which would predict similar effects on effort of partner contribution and of the randomly assigned multiplier. In

our data only the former affects effort levels.

The right half of table 2.1 repeats the analysis for the UPFRONT scenario. Columns 5 and 6 show a substantial effort gap between EQUAL and each NON-EQUAL contracts, and an increase in the gap as we go from stage I to stage II. In particular, PROPORTION contracts are associated with an effort increase of 28 (37) percentage points in stage I (stage II) relative to EQUAL. Given that the stage I (stage II) effort gap was 8 (10) percentage points in the IMPOSED scenario, over 70% of the differences in contract performance appear to be driven by factors other than the incentive strength of the contracts. Further, each VESTING contract is associated with lower effort relative to PROPORTION contract. However, these differences are at most 9.6 percentage points (Wald tests, $p = 0.017$ and $p = 0.000$). There are no significant differences between the two VESTING contracts ($p = 0.255$). Columns 7 and 8 suggest that some of the changes in effort are, again, a result of the subjects reacting to observed partner behavior, and not to differences in incentive strength. Column 8 shows that the effect of own stage I multiplier on stage II effort is not statistically significant ($p = 0.529$) whereas the effect of partner contribution is statistically significant ($p = 0.000$).

Summing up our results so far, EQUAL contracts are associated with uniformly lower effort levels compared to each NON-EQUAL contract. However, over 70% of the effort gap is driven by factors other than the incentive strength of the contract. Even with exogenously imposed contracts effort differences are driven in part by reactions to partner behavior, and not by the strength of incentives alone.

2.4.4 Discussion

While our results are consistent with the conventional wisdom that equal splits are associated with low value generation, our data suggest that this is not driven by the differences in incentive strength. If incentive strength drives the differences

Table 2.1: Effects of contract form on effort

	Treatment: IMPOSED				**Treatment: UPFRONT**			
Dep. Var:	stage I effort	stage II effort	stage II effort	stage II effort	stage I effort	stage II effort	stage II effort	stage II effort
EQUAL	baseline	baseline	baseline	baseline	baseline	baseline	baseline	baseline
THRESH VEST					21.655***	29.391***	25.566***	27.052***
					(4.855)	(5.156)	(4.649)	(5.010)
DIFF VEST					20.139***	27.286***	25.258***	25.687***
					(4.398)	(5.337)	(4.903)	(5.132)
PROPORTION	8.004*	9.514**	7.391*	8.598**	27.803***	36.871***	31.850***	33.634***
	(4.375)	(4.295)	(4.031)	(4.266)	(5.018)	(6.270)	(5.743)	(6.021)
V_1			0.695***				0.821***	
			(0.164)				(0.117)	
Own stage I multiplier				0.358				0.855
				(2.221)				(1.357)
Partner stage I contribution				0.639***				0.742***
				(0.209)				(0.130)
Constant	43.827***	38.930***	31.271***	36.048***	28.384 ***	14.675	8.357	11.821
	(8.706)	(9.004)	(9.168)	(9.487)	(9.407)	(9.757)	(9.179)	(9.657)
Observations	400	400	400	400	432	432	432	432
Subjects	50	50	50	50	54	54	54	54
					Tests of linear combinations of coefficients			
THRESH VEST−DIFF VEST					1.516	2.106	0.309	1.365
					(2.303)	(1.850)	(1.960)	(1.918)
THRESH VEST−PROPORTION					-6.147**	-7.480**	-6.284**	-6.582**
					(3.050)	(3.122)	(2.677)	(2.626)
DIFF VEST−PROPORTION					-7.664***	-9.585***	-6.593**	-7.947***
					(2.560)	(2.598)	(3.129)	(3.115)

Note. Dependent variable is stage I effort (columns 1 and 5) and stage II effort (columns 2-4 and 6-8). Regression coefficients are obtained using random effects regression, standard errors clustered at subject level. Controls: age, gender, experimental period.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

in effort, we should observe robust differences in contract performance even when contracts are imposed exogenously rather than being selected endogenously by the team. However, we see the opposite, that 75% of the effort gap between contracts disappears when contracts are imposed externally. In the latter scenario selection is not possible, suggesting that the effort gap between contracts is driven by the individuals with socially desirable behaviors selecting into non-equal contracts and vice versa.

Further, if subjects respond to incentive strength alone (as standard theory would predict), we should observe similar reactions to partner-driven and exogenous changes to the marginal return for investing effort into the startup. However, we again see substantial deviations from theory predictions. Effort levels do not change in response to exogenous changes in productivity but do change in response to partner effort, suggesting that initial effort can be a salient signal that drives (or reduces) cooperative behaviors in the team. In the next section we show that not only effort levels, but also contract choices can be affected by initial effort signals when effort information is available to the team prior to contract selection.

In addition to the effort gap between equal and non-equal contracts our data reveal some differences between the non-equal contracts. However, these differences are substantially smaller, relative to the equal/non-equal gap. This, again, is consistent with the incentive strength being a secondary factor in our data. If incentive strength was the main driving force, we should observe robust performance differences between proportional contracts and vesting contracts because vesting contracts impose only a mild penalty for free-riding. However, effort and value differences between vesting and proportional contracts are small. In section 6 we show that this is primarily because vesting contracts attract fewer undesirable founder types, relative to equal contracts.

Selection patterns similar to ours have been observed in the experimental eco-

nomics literature (*Gürerk et al.*, 2006; *Sutter et al.*, 2010; *Dal Bó et al.*, 2010). This literature shows that individuals with undesirable behaviors are frequently opposed to high-powered incentive regimes. However, much of the literature focuses on examining behavior in one-shot interactions characterized by a conflict between what is socially efficient and individually optimal. Our study is different in that it presents teams with a more contextualized environment (involving risky and partner-dependent returns to effort investment) that is more reflective of collaborative work in startups. Further, the contracting process itself is designed to reproduce the contracting dynamics in startups with a range of available contracting options from equal to fully contribution-proportional. Such contractual division rules have not been examined in the literature, which has mainly focused on voting-based reward allocation and voluntary punishment of free-riders (*Gürerk et al.*, 2006; *Cappelen et al.*, 2007; *Sutter et al.*, 2010; *Dal Bó et al.*, 2010).

Our results so far suggest that contractual offers may signal something about the personality type of the individual when contracting happens upfront. Many entrepreneurial teams, however, delay contracting until at least some work is done (*Wasserman*, 2012; *Hellmann and Wasserman*, 2016). In that case founders can observe each other's collaborative behaviors, which can provide another signal into the personality of the partner, prior to contracting. In section 5 we investigate the consequences of this additional signal.

2.5 Delayed contracting

In this section we examine a scenario in which equity contracting is delayed until after stage I. The sequence of events is similar to the UPFRONT contracting treatment, however the order of the stage I contribution phase and the negotiation phase is reversed.

2.5.1 Model parameters and equilibrium predictions

As in the UPFRONT treatment, each subject is endowed with 10 units of effort to be allocated between the risky startup account and the safe private account in each of the two contribution stages. As previously, the effort allocated to the startup account is multiplied by 0.5, 1 or 2 with probabilities 0.25, 0.5 and 0.25. The contract parameters are as follows:

$$\sigma_i^{EQUAL} = 0.5$$

$$\sigma_i^{THRESH\ VESTING} = \begin{cases} 0.3 & \text{if } \{c_{i2} < c^{THRESH} \wedge c_{j2} \geq c^{THRESH}\} \\ 0.7 & \text{if } \{c_{i2} \geq c^{THRESH} \wedge c_{j2} < c^{THRESH}\} \\ 0.5 & \text{otherwise} \end{cases}$$

$$\sigma_i^{DIFF\ VESTING} = \begin{cases} 0.3 & \text{if } \{c_{j2} - c_{i2} \geq c^{DIFF}\} \\ 0.7 & \text{if } \{c_{i2} - c_{j2} \geq c^{DIFF}\} \\ 0.5 & \text{otherwise} \end{cases}$$

$$\sigma_i^{PROPORTION} = \frac{c_{i2}}{c_{i2} + c_{j2}}$$

Notice that because our investigation focuses on forward-looking, ex ante symmetric contracting, the allocation of shares in DELAYED NON-EQUAL contracts is based on stage II contributions and is independent of stage I contributions. The equilibrium structure is similar to the UPFRONT scenario. Different contracts feature different off-equilibrium path predictions for stage II, but identical (full effort) predictions for stage I. Complete characterization of the equilibrium is relegated to the supplementary materials.

2.5.2 Experimental results

Aggregate effort and value levels (averaged over all contracts) are similar in the UPFRONT and the DELAYED treatments. On average, subjects invest 67 (70) percent of their effort endowment and their team value V is 278 (273) points in the UPFRONT (DELAYED) treatments.[9]

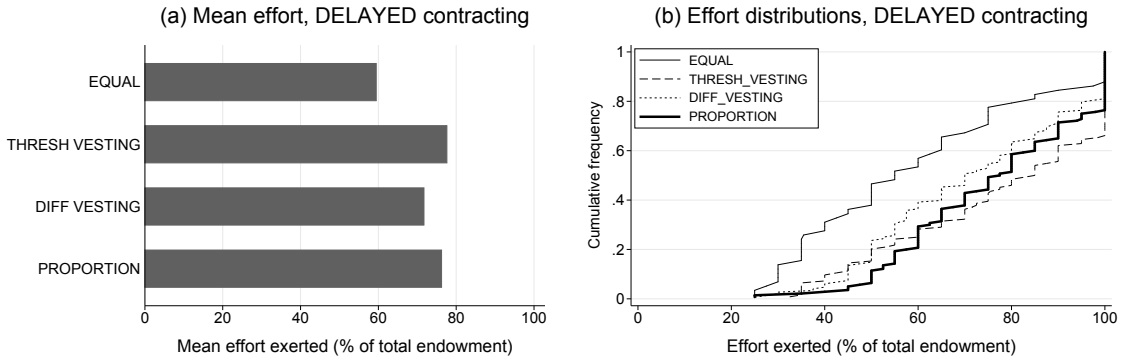
However, there appears to be an interactive effect of contract form and contracting time on effort (see fig. 2.2). In particular, the effort gap between EQUAL and NON-EQUAL contracts, as well as the gap among the NON-EQUAL contracts both shrink substantially in the DELAYED treatment relative to the UPFRONT scenario. When contracting is delayed, the gap between EQUAL and THRESH VESTING (DIFF VESTING) is 17.8 (12.3) percentage points. Further, THRESH VESTING exhibit effort levels on par with PROPORTION, while both THRESH VESTING and PROPORTION perform better than DIFF VESTING with the difference of 4.53 and 5.90 percentage points, respectively.

Regression analysis

Table 2.2 examines the effects of contracts on effort levels more formally, using random effects regressions. Column 1 shows that none of the stage I effort differences between contracts are statistically significant. This result is strikingly different from the UPFRONT scenario, in which we saw substantial stage I effort differences between contracts. Column 2 shows that both VESTING and PROPORTION contracts are associated with increased stage II effort relative to EQUAL contracts (all $p < 0.01$).

[9]Random effects regressions confirm that there are no significant effort or value differences between pooled UPFRONT and pooled DELAYED treatments ($p > 0.1$). However, teams split their effort endowments somewhat differently between stage I and stage II in UPFRONT and in DELAYED. Given the multiplicative structure of the startup value function ($V = V_1 \times V_2$) the efficient allocation of a fixed amount of effort (from the team perspective) would be to split effort evenly between the stages. Such even allocations of effort are observed in all UPFRONT contracting scenarios. In contrast, in all DELAYED NON-EQUAL scenarios subjects increase effort by 15 to 20% as they go from stage I to stage II.

Figure 2.2: Mean effort and effort distributions in the DELAYED treatment



However, these effort differences shrink by about 60 to 70 percent, relative to the UPFRONT scenario suggesting that contract form and contracting time have an interactive effect on effort. Column 3 of table 2.2 shows that the effect of V_1 on effort is significant ($p < 0.01$), but the strength of the effect is, again, reduced relative to the UPFRONT scenario (UPFRONT: $\beta = 0.821(0.117)$, DELAYED: $\beta = 0.369(0.088)$). Column 4 suggests that this is driven primarily by a drop in the effect of stage I partner contribution on subsequent effort, both in terms of magnitude and statistical significance (UPFRONT: $\beta = 0.742(0.130)$, DELAYED: $\beta = 0.197(0.129)$).

In sum, in the DELAYED scenario more egalitarian regimes perform better while the contribution-proportional regime performs slightly worse relative to the UPFRONT scenario. Further, partners' stage II effort is less sensitive to mutual stage I effort levels. Both these effects may be driven by the availability of effort information prior to contracting, allowing founders to identify free-riders early on and to reduce the appeal of free riding by choosing NON-EQUAL contracts. This behavior will be examined next.

Table 2.2: Effects of contract form on effort in the DELAYED treatment

	Treatment: DELAYED	Treatment: DELAYED	Treatment: DELAYED	Treatment: DELAYED
Dep. Var:	stage I effort	stage II effort	stage II effort	stage II effort
EQUAL	baseline	baseline	baseline	baseline
THRESH VESTING	2.508	12.501***	13.128***	12.479***
	(3.348)	(4.123)	(4.116)	(4.163)
DIFF VESTING	0.272	11.154***	11.664***	11.175***
	(2.867)	(3.536)	(3.523)	(3.549)
PROPORTION	0.804	16.611***	17.318***	16.680***
	(3.266)	(3.965)	(4.002)	(3.999)
V_1			0.369***	
			(0.088)	
Own stage I multiplier				0.356
				(1.022)
Partner stage I contribution				0.197
				(0.129)
Constant	36.624***	40.602***	37.131***	39.497***
	(12.248)	(12.032)	(11.565)	(12.144)
Observations	470	470	470	470
Subjects	56	56	56	56
	Tests of linear combinations of coefficients			
THRESH VESTING−	2.235	1.347	1.465	1.304
DIFF VESTING	(2.041)	(2.245)	(2.262)	(2.245)
THRESH VESTING −	1.704	-4.110*	-4.189*	-4.201*
PROPORTION	(2.166)	(2.273)	(2.287)	(2.377)
DIFF VESTING −	-0.531	-5.457**	-5.654**	-5.505**
PROPORTION	(1.977)	(2.367)	(2.303)	(2.296)

Note. Dependent variable is stage I effort (columns 1) and stage II effort (columns 2-4). Regression coefficients are obtained using random effects regression, standard errors clustered at subject level. Two observations were removed because the team was not able to agree on a contract. Five of the six experiment sessions involved 8 experimental periods, one session involved 10 experimental periods (programming error). Qualitative results are not sensitive to omitting that session. Controls: age, gender, experimental period.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

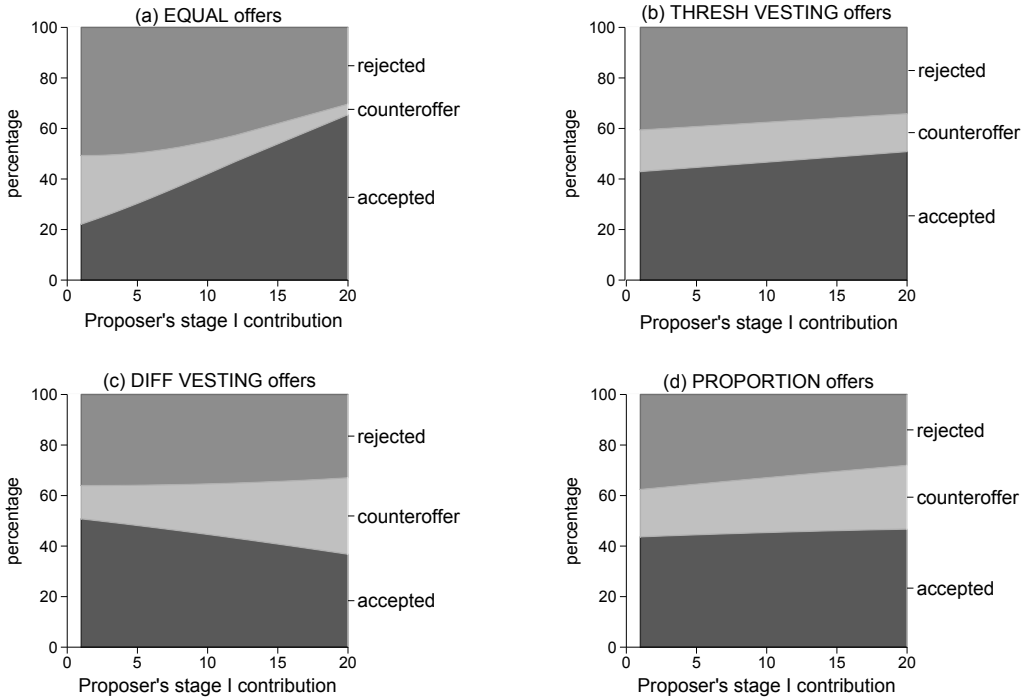
Negotiation dynamics and contract choices in DELAYED

Similarly to the UPFRONT scenario, each contract is chosen by a non-trivial share of the teams in each experimental round, and no contract dominates the contracting decisions in the DELAYED treatment. However, an examination of the negotiation dynamics reveals some important differences between the UPFRONT and DELAYED treatments. In particular, in the DELAYED scenario teams negotiate longer (UPFRONT: 23.47 seconds, DELAYED: 34.99 seconds, random effects regression: $p < 0.001$) and exchange significantly more offers before agreeing on a contract (UPFRONT: 2.09 offers, DELAYED: 3.04 offers, $p < 0.001$). These results suggest that individuals are more persistent in pursuing their contract preferences when contracting is delayed.

Further, our data suggest that the increased intensity of the negotiations in the DELAYED scenario is driven primarily by the teams arguing about choosing (or not) EQUAL division contracts. On average, when EQUAL contracts are mentioned during the negotiations teams spend more time negotiating (49.15 seconds vs. 27.87 seconds, random effects regression: $p < 0.001$) and exchange more offers (4.73 offers vs. 2.18 offers, $p < 0.001$), relative to the teams that do not consider EQUAL contracts.

In addition to increased negotiation intensity in the DELAYED scenario, the probability of the partner accepting an EQUAL offer is positively correlated with the proposer's stage I contribution level ($\rho = 0.20,\ p = 0.056$). Indeed, EQUAL contract proposers can convince their partners to agree to an equal split only if they show evidence of high effort, as shown in figure 2.3 (see Appendix B.2 for estimation details). Specifically, if the proposer contributes nothing to the startup in stage I, her EQUAL offer will be accepted with 20% probability. However, if the proposer contributes the maximum possible value, 20 points, then her offer will be accepted with 60% probability. On average, the odds of an EQUAL contract being accepted

Figure 2.3: Response to contract proposals, as a function of proposer's stage I contribution (DELAYED)



Note. Predictive margins of response to contract offers are displayed. Predictions are obtained using Multinomial Logit regressions of response (0: reject, 1: counteroffer, 2: accept) on the proposer's contribution level. Standard errors are clustered at subject level. For the regression specification and detailed estimation results see Appendix B.2.

increase by 11% with each contribution point ($p = 0.042$). In contrast, the probability of acceptance is near-constant at 40 to 50 percent for each NON-EQUAL contract, as illustrated in figures 2.3(b)-(d). In fact, the relationship between the proposer's contribution level in stage I and her partner's response is not statistically significant for any NON-EQUAL split offer ($p > 0.268$).[10]

In sum, the analysis of the negotiation dynamics suggests that with delayed contracting initial effort is an important signal used by teams to decide on the contracts. The availability of effort information affects equal contract proposers who are scrutinized more closely by their partners prior to agreeing to an equal split offer.

[10]Figure 2.3 uses Multinomial Logit predictions with the response to contract offer as dependent variable (0: accept, 1: counteroffer, 2: reject). For robustness we repeat the analysis using random effects Logit regressions with a binary dependent variable (0: accept, 1: not accept) and find similar results.

2.5.3 Discussion

When contracting happens prior to the start of collaboration equal division contracts are associated with poor performance, mainly due to self-selection of free-riders into equal contracts. But, when contracting is delayed the effort and value gap between equal and non-equal contracts narrows by more than 60% relative to the upfront scenario. Further, the performance gap among the non-equal contracts, too shrinks to a minimum; in fact effort and value levels in the different non-equal contracts become statistically indistinguishable.

Our investigation of the negotiation dynamics suggests that the narrowing of the contract performance gap is the result of the change in the information available to the team when they select contracts. With upfront contracting, the contract offers are the only signal available to the teams and there are few barriers for free-riders to select into equal division contracts. Further, free-riding (and cooperative) behaviors are reinforced as team partners reciprocate to each other's initial contributions levels. Together, these effects lead to a robust performance gap between equal and non-equal contracts. However, with delayed contracting initial effort is another signal into the personal characteristics of the proposer. Free-riders are revealed early on by their partners, who can refuse equal contracts if they observe low initial effort signals. Further, robust contributors are also revealed early on and others are willing to sign equal contracts with them.

The positive effect of delayed contracting on equal contract performance, while being conjectured in the entrepreneurial contracting literature, has not been validated by data (*Hellmann and Wasserman*, 2016; *Jared*, 2016). More generally, the entrepreneurship literature is hesitant to recommend delayed contracting listing two undesirable features of postponing the negotiations. First, not knowing how their efforts will be rewarded, founders may be reluctant to invest effort in the pre-contracting stages. Second, the value of the business (often) increases over time raising the stakes

for the founders, which may lead to increased conflict potential and extend the negotiations (*Wasserman*, 2012). Our data confirm both these features of delayed contracting. However, our results show that these effects are dominated by the additional effort information exchanged prior to the negotiations, allowing founders to match contracts to personality type. For founders who are concerned with free-riding in their teams but skeptical about performance-based contracts, these results suggest that delaying the contracting can improve performance.

In sum, in the delayed contracting scenario teams' contracting decisions are driven at least partly by the initial effort signals, with the consequence that undesirable founder types are no longer able to self-select into equal contracts. However, the new information available to teams prior to contracting may have other, more indirect effects on behavior of both undesirable and desirable founder types. The next section examines these effects more closely.

2.6 Characterization of types' preferences and behaviors

Our results so far indicate that the effort gap between contracts is driven primarily by the differences in personal characteristics of individuals who select these contracts. Further, the information available to the team prior to contracting matters for contract performance. In the UPFRONT scenario revealed negotiation preferences is the only information available to negotiators, but in the DELAYED scenario teammates have additional information to incorporate into their expectations for future performance. In this section we examine these selection dynamics from a new angle, introducing a taxonomy of personality types who signal their type by the contract form they favor in negotiations. Specifically, our data suggest three types—low contributors, conditional contributors and high contributors—and these behave differently in negotiations and perform differently even under identical contracting regimes and when faced with different partner behaviors.

The examination of type behaviors addresses three open questions in our investigation. First, we have seen that behaviors in our data are driven not solely by self-interest, or at least not in ways predicted by standard incentive theory, so it is useful to characterize the relevant drivers of behavior more explicitly. Second, we have shown that EQUAL contracts attract free-riders. However, it is also useful to examine the differences between those who favor VESTING and those who favor PROPORTION contracts. Third, effort information available to the team prior to contracting has been shown to affect contract performance. Given that each personality type may put different weights on different outcomes, it may be worthwhile to examine how this additional effort information interacts with the types' preference structures.

We next discuss type assignment, the preference structure of each type and type behaviors in each treatment. The description of our estimation methodology, the robustness analyses and a more detailed discussion of the estimation results are left to the supplementary materials.

2.6.1 Type assignment and types' preferences

While there are many procedures to divide the subject population into types, we use the contract offers subjects accept and reject in negotiations. These negotiation data give a more nuanced window into personality types than the final contracts, because they reflect individual preferences for different division rules (and not team consensus). In our case, the availability of three contractual alternatives lends itself to a three-type taxonomy (low contributors, conditional contributors and high contributors) with each type preferring one of the three contract forms (EQUAL, VESTING, PROPORTION). The label choices for the types will become clear below.[11]

[11]We identify low (conditional, high) contributors as subjects who prefer EQUAL (VESTING, PROPORTION) contracts to other contract forms in the initial three experimental rounds. We pool THRESH VESTING and DIFF VESTING into one category because we do not find substantial differences in behaviors between subjects who prefer one of these contracts.

Table 2.3: Types' preference structure.

	Relevance for type		
Factor	*Low contributor*	*Conditional contributor*	*High contributor*
Own profit	Yes (marginally sign.)	Yes	Yes
Profit differences within team	No	Yes	No
Effort differences within team	No	Yes	No
Exerting less effort than partner	Yes	No	No

Note. Each factor is computed in terms of expected values. Profit differences within team are computed in relative terms, based on the expected deviation of the share allocated to player i and the 50% norm (*Bolton and Ockenfels*, 2000). Effort differences are computed as expected absolute difference between own effort and partner effort. "Yes" indicates that the corresponding factor is associated with a statistically significant utility coefficient in the type's utility function ($p < 0.01$). "No" indicates that the factor is not associated with a statistically significant utility coefficient in the type's utility function ($p > 0.1$). For detailed results see the supplementary documents.

To allow insight into the drivers of type behaviors we use Conditional Logit analysis (*McFadden*, 1973). The range of Conditional Logit uses is extensive, but the closest application to ours is the analysis of distributional preferences in the experimental economics literature (see *Frey and Meier*, 2004; *Bardsley and Moffatt*, 2007; *Cappelen et al.*, 2007; *Moffatt*, 2016). In these studies, and in ours, Conditional Logit is used to characterize the preferences of a population by estimating the coefficients in utility models that account for self-interest (profit maximization) and a range of nonself-interest factors. The functional forms of these models are chosen based on their ability to explain the data, both in terms of adding intuition and their econometric fit.

Results

Our estimation results indicate that all types are at least partly concerned with own profit maximization. However, the extent to which other factors (not related to narrow self-interest) affect their decisions differs by type. In particular, low contributors are primarily driven by the desire to work less yet share equally in any profits

generated. Further, they put the lowest weight on their own profits relative to the remaining types; in fact the utility coefficient on their own profit is only marginally significant. Indeed, our models show that low contributors are the only type who will tolerate lower earnings if they can invest less effort than their partners. Conditional contributors care more about own profits than low contributors, but are also concerned with effort and payoff of their partners. They will tolerate lower earnings if they can avoid discrepancies in both effort and profits in the team. High contributors are not concerned with anything other than their own payoffs. None of the coefficients on other, nonself-interest factors are statistically significant for them. Table 5 summarizes these results.[12]

2.6.2 Type behaviors

We next contrast type behaviors in the UPFRONT and DELAYED treatments. We use effort and contracting data from experimental periods 4-8 for these comparisons (Periods 1-3 are excluded because they were used to assign subjects to types).

UPFRONT treatment

Figure 2.4(a) shows the contracts preferred by each type.[13] The data reveal that low contributors prefer EQUAL and PROPORTION contracts to VESTING contracts. Conditional contributors are not entirely opposed to EQUAL contracts and PROPORTION contracts, but typically lead with VESTING offers. High contributors express a preference for PROPORTION contracts most of the time. They never

[12]In this analysis, the utility coefficients are estimated using the UPFRONT data. The reason for omitting the DELAYED data in the utility analysis is the interaction of effort signals and revealed contract preferences in the DELAYED scenario. Detailed description of type assignment for both UPFRONT and DELAYED treatments and the characterization of the utility functions is relegated to the supplementary documents.

[13]We measure contract preferences of the types by tracking the initial contract offers they make in the negotiations. If a subject makes no offers in a given round we use the contract he/she accepts in the negotiations. For robustness we replicate the analysis using rejection and acceptance rates of contracts, for each type. These robustness checks yield similar results.

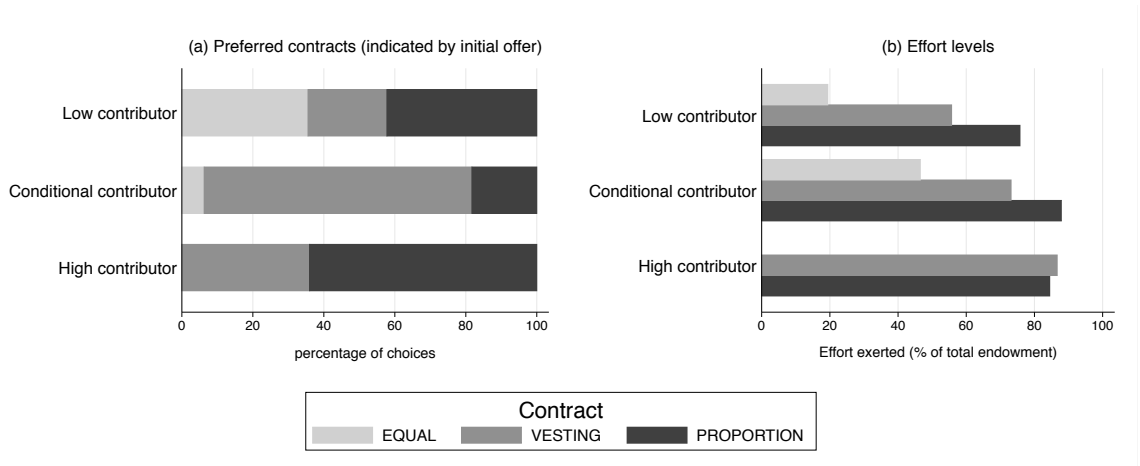


Figure 2.4: Contract preferences and effort levels by personality type (UPFRONT treatment)

Note. Contract preferences show the shares of first contract proposals for each type in the UPFRONT treatment (If a subject did not propose any contracts in a given round, he/she is assigned the first contract he/she accepts). Effort comparisons show total (stage I + stage II) effort as a percentage of endowment. The bar for high contributors in EQUAL is omitted in panel (b) because high contributors never select EQUAL contracts.

lead with EQUAL contracts, but sometimes accept and offer VESTING contracts.

The differences in revealed contract preferences of the types were confirmed in Probit regression analysis. Conditional contributors exhibit a stronger preference for VESTING and a weaker preference for EQUAL contracts relative to low contributors, whereas high contributors exhibit a stronger preference for PROPORTION contracts relative to conditional contributors (all $p < 0.05$, see Appendix B for the estimation results). These preferences align with intuition. Further, the types differ not only in the contracts they favor, but also in the contracts they end up selecting, confirming that the personality mix differs significantly between contracts.[14]

We now turn to the differences in effort levels between the types. Figure 2.4(b) reveals that both low and conditional contributors' effort levels are sensitive to the chosen contract form, whereas high contributors are indifferent to the contract form.

[14]The result that low contributors prefer PROPORTION to VESTING contracts may appear surprising. However, recall that low contributors care mainly about working less than their partners. This can be done most easily in PROPORTION contracts, in which partner effort is highest among all contracts.

Further, low contributors exhibit lower effort levels in each contract form relative to the remaining groups. To examine these differences more formally we regress effort levels on the personality type dummy variable and the chosen contract (for estimation results see Appendix B). Indeed, our regression results confirm that low contributors exhibit lower effort relative to both conditional and high types, even after controlling for the contract. Further, comparisons of within-type effort levels suggest that both low and conditional contributors are more sensitive to the contract form than high contributors and that conditional types are the only ones who adjust their stage II effort levels based on observed stage I partner effort (for estimation results see Appendix B).[15]

DELAYED treatment

In the UPFRONT treatment individuals signal their type by the contract offers they make and accept in the negotiations. However, in the DELAYED scenario initial effort is another signal that may affect how types behave in the negotiations. Therefore, rather than identifying types based on their initial contract offers (as we did in the UPFRONT scenario), we assign types based on type similarity scores that are computed for each subject based on their contracting and stage II effort decisions.[16]

Comparisons of type behaviors reveal that the availability of additional effort information changes the contract forms each type prefers in the negotiations and also affects the effort levels they exhibit in contracts. In particular, we have seen that low

[15] We also examine whether types differ in the extent to which they act according to the equilibrium predictions. We examine differences in type behavior when equilibrium analysis predicts full effort provision and when it predicts zero effort provision in stage II and find that both low and conditional contributors deviate from the equilibrium predictions more often than not, and by large amounts, whereas high contributors act in accordance with equilibrium predictions in 73% of the cases.

[16] To account for the differences in the initial partner efforts subjects see prior to contracting, we use these data to compute the utility that each subject would enjoy conditional on being low, conditional or high type, given the contracting and stage II effort decisions they make. Using subjects' decisions in periods 1 through 8 we then compute the posterior probabilities for each subject of being the low, conditional or high contributor type and assign each subject to the type with the highest posterior probability. Our estimation procedure and estimation results are described in detail in the supplementary materials.

contributors want to maximize their own payoffs, subject to low effort expenditure. In the UPFRONT treatment, these preferences led to low contributors self-selecting into EQUAL contracts. However, in the DELAYED treatment free-riding intent is revealed early on, with the consequence that low contributors are forced to either increase initial effort (so that others sign equal contracts with them) or choose NON-EQUAL contracts. Our results show that both these behaviors indeed occur in the DELAYED scenario (Detailed results are relegated to the supplementary documents).

When contracting happens prior to collaborating, conditional contributors are not entirely opposed to EQUAL contracts, but prefer VESTING contracts in which they can avoid excessive free-riding and also reduce profit discrepancies in the team. However, because they are facing more desirable behaviors in DELAYED EQUAL contracts, they become more receptive to EQUAL offers, particularly when they see high initial effort.

High contributors are strongly opposed to EQUAL contracts in the UPFRONT treatment. They exert maximum effort and prefer the strongest incentive scheme to induce their partners to do the same. However, if they observe high initial efforts they may be less inclined to insist on PROPORTION contracts, particularly if they believe that initial effort is predictive of future behavior. This should lead to the high contributors becoming more receptive to both EQUAL and VESTING contracts, relative to the UPFRONT scenario. Indeed, our data show that high contributors sometimes (though still less frequently than other types) choose EQUAL contracts, and also continue contributing near-maximum effort in DELAYED EQUAL.

In the aggregate, in the DELAYED scenario low contributors select into EQUAL contracts at a lower rate, while conditional and high contributors select into EQUAL contracts at a higher rate, relative to the UPFRONT scenario. This results in a more even personality mix in each contract, reducing effort and value differences between contracts.

2.6.3 Discussion

Our data reveal the presence of three behavioral types (low, conditional and high contributors) that differ in their motives and behaviors. While all three types prefer more profit to less, the preference structure for two of the three types is more complex and features tradeoffs of profit for other considerations. Specifically, low contributors do not want to work more than their partners, and will tolerate lower payoffs if they can work less than their partner. Conditional contributors dislike discrepancies in both effort and payoffs within the team. High contributors are driven by self-interest alone, in line with the preferences frequently assumed in standard economic analyses.

In the upfront scenario contractual offers and responses to these offers are the only signals that founders have to work with. Low contributors signal their type by offering and accepting equal contracts, because these allow them to work less yet share equally in any profits generated. If low types end up in non-equal contracts they are not indifferent to incentive strength, but still exert lower effort relative to any other group in that contract. Conditional contributors prefer vesting contracts because these reduce pay inequalities, relative to proportional splits, and at the same time limit free-riding behaviors, relative to equal splits. Conditional contributors' effort levels differ with the contract, however a significant part of the within-type effort difference is driven by their attempts to match partner effort and not by the incentive strength of the contract. High contributors are not concerned with anything other than their own payoffs. They exert near-maximum effort and prefer proportional contracts to vesting, because the former hold their partners fully accountable for their actions. Further, high contributors strictly avoid equal contracts.

When contracting happens upfront, these behaviors lead to the low contributors being overrepresented in equal contracts, and the other types being overrepresented in non-equal contracts. That is, equal contracts are bad for team performance, not primarily because of their incentive strength but because of the founder types that self-

select into them. But, in the delayed contracting scenario, founders have additional information to work with: the initial contribution of their partner. Low contributors are revealed and others do not want to sign equal contracts with them. Further, robust contributors are also revealed which reduces others' reluctance to sign equal contracts with them. Together, these behaviors result in low contributors no longer being over-represented in equal contracts. More generally, since it is founder type rather than the contract type (strength of incentives) that primarily impacts behaviors, with a stronger signal of type the contract form becomes less important leading to a more even distribution of types over contracts and to smaller effort and value differences between contracts, relative to the upfront scenario.

Taken together, these results add texture to the signaling and selection dynamics described in the previous sections. Different personality types have different desires, and they pursue these desires consistently, in each contracting regime. However, the availability of effort information that is not tied to contract means that in the delayed scenario different contracts can attract their attention. For startup contracting in practice these results suggest that the presence of undesirable personality types in the team can be best handled by delaying the negotiations until further downstream in the entrepreneurial innovation process.

2.7 Concluding remarks

This is the first experimental test to our knowledge of the relationship between contract form and contracting time and effort and value generation in startups. Our results confirm the conventional finding that equal splits are poor choices, but suggest that this is driven not by the incentive differences between contracts, but mainly by the differences in personality types. Equal splits are proposed and embraced by the least desirable personality types who prefer working less than their partners even when this harms their profits. Contribution-dependent contracts attract high contributors

who invest high effort and prefer the strongest incentive for the partner to do the same.

We also find a smaller contribution gap, between partially performance-dependent (vesting) contracts and fully performance-dependent contracts. This is, again, driven by the differences in personality type. Vesting contracts are preferred by individuals who care not only about their own payoffs, but also dislike disparities in both, effort levels exerted by the team and payoff levels. In contrast, individuals who choose fully performance-dependent contracts are guided entirely by self-interest.

When contracting happens upfront, individuals can often select contracts that align with their preferences. This generates a substantial effort gap between equal contracts (dominated by undesirable personality types) and non-equal contracts (dominated by desirable types). As a result, teams choosing equal contracts generate only half of the value relative to the teams choosing non-equal contracts. However, when contracting is delayed the effort and value gap between equal and non-equal contracts shrinks by about 60 percent and the differences between the non-equal contracts disappear completely.

The narrowing of the performance gap between contracts is the result of the change in the signaling and selection dynamics. When contracting happens upfront revealed contract preferences are the only signal available to the team, but in the delayed scenario there is an additional (costly) effort signal that parties use to indicate who they are and how they will behave in the future. In the presence of this additional signal, equal contracts are accepted only when equal contract proposers demonstrate that they are worthy of an equal split, by exerting high initial effort.

Our findings have several implications for startup investors. Our results confirm the conventional wisdom that investors should avoid startups with equal equity splits between founders. However, equal contracts chosen further downstream in the entrepreneurial process are markers of a more desirable personality mix, relative to equal

contracts selected early on. Hence, information about when and how the contract was chosen is as important as the contract form and should be included into the investors' due diligence process.

Our findings also have implications for startup founders. Because personality characteristics are the primary driver of behavior founders should pay as much (or more) attention to personality type as they do to contract form. For a given mix of personality types, however coalesced, the contract form can make a difference. However, especially in the early stages founders may have neither the flexibility to change the composition of their teams, nor the ability to include stringent terms into their equity contracts. In such situations, delaying the contracting can improve performance.

Our investigation focuses on founder teams formed by peers with few differences in prior founder experience, who can (at least initially) be expected to add similar value if they decide to participate in the venture. Such teams are a common, but not the only form of early-stage startups. An important next step is to examine the effects of different contracting regimes in teams that are more asymmetric and differ in seniority (founder/advisor or inventor/first-employee teams) and expertise (technical-developer/marketer teams).

CHAPTER III

Entrepreneurial Market Research

3.1 Introduction

It is a common feature of new product and venture-capital portfolio management in industry that decision-makers trade off risk and return when choosing new additions to their portfolios. Surveys of practice reveal that the solicitation of the risk and return characteristics of potential projects can take a variety of forms, but it is common to assess these two dimensions of a potential investment separately (*Cooper et al.*, 2006). This is because high risk high return investment opportunities are perceived differently by portfolio managers than low risk low return opportunities, even if their "expected value" is similar.

The assessment of the risk and return characteristics for any potential investment can take a variety of forms. For example, structured questions about market size, the level of competition, the uniqueness of the proposed product, and other characteristics may be aggregated into a summary value measure. Separately, questions about translational technology risk, regulatory risk, supply chain risk, commercialization risk, and other sources of uncertainty are aggregated into a separate summary risk measure. Portfolio managers then consider both of these dimensions when comparing potential investments.

Some methods used in practice invoke a structured way to trade off risk and return.

For example the ECV (Expected Commercial Value) method simply multiplies the value estimate by a probability of success. Other methods present the value and risk assessment separately and allow the portfolio managers to judgmentally trade them off. For example, risk return diagrams or bubble diagrams plot the various projects on the two-dimensional (risk-return) plane to give decision-makers a visual representation of those characteristics of potential investments.

The common feature among all these methods is the solicitation of a value metric and a risk metric for each potential investment and the consideration of these two when choosing additions to the portfolio. This is also a common feature of entrepreneurial teams investigations into potential markets, wherein semi-structured interviews of key informants solicit impressions of the potential value, and the risks, for entering a given market, because these features will influence how potential investors value their venture.

Our objective is to inform this investigative process in the context of a startup team which has developed a new material or technology that could potentially find voice in a range of consumer and industrial applications, and the team wishes to identify the most promising market direction in which to take their invention. We assume that the team operates with finite time and budget, restricting the number of markets that can be explored, and/or the depth in which they can explore any one of them. The team explores a market by interviewing key informants in that market and soliciting from them the potential value of their invention and the risks they would face achieving that potential value in that market.

The more interviews the team can conduct in any one market the better their estimates will become of the risks and potential returns for entering that market. However, with finite resources, the team faces a classic exploration versus exploitation trade-off in their allocation of time and budget. This means, they need to decide carefully how many markets to explore, in what sequence and in what depth. In this

paper we propose and test a range of search strategies to answer these questions.

Perhaps the most extensive literature studying the exploration-exploitation trade-off is the Multi-arm bandit literature (For a summary of fundamental bandit results and classic search policies see, for example, *Auer et al.*, 2002; *Gittins et al.*, 2011; *Bubeck et al.*, 2012). Leveraging this robust literature we study the performance of the classic bandit search strategies and develop several new ones for three bandit variants, accounting for the key features of the entrepreneurial market identification problem.

The first variant is an extension of the classic Bernoulli bandit model, with the additional consideration of sampling variance in the objective function. The Bernoulli model is somewhat limited in that it can incorporate only sampling risk, and not the inherent risk differences that are independent of the value of the arm (as the variance in an arm's performance is pre-ordained by the arm's probability of success). To incorporate inherent risk differences between the markets we introduce a second model, adapted from the Normal bandit model. In the Normal model our objective will be to discover the alternative with the maximum mean-variance score, defined as the convex combination of the mean and the negative value of the variance with pre-specified weights on the mean and the variance components.

While the Normal model allows a richer characterization of risk relative to the Bernoulli model, the informational content of each signal is (similarly to the Bernoulli model) collapsed into one scalar value. That is, neither the Bernoulli nor the Normal models can separate between the information about market value and about market risk obtained during the market research. To examine policy performance in a setting with a richer signal content, we develop a third model that we will refer to as the Risk-return model. The Risk-return model retains the customary weights on the risk and return metrics, but measures risk via a second (risk-related) signal obtained in each stage from the sampled alternative

In addition to the described risk-return trade-offs, our investigation accounts for three additional features of the entrepreneurial market identification problem: the high cost of obtaining market information, the postponement of rewards until the end of the search phase, and the requirement that the recommended policies can be implemented by a human decision-maker.

The cost of conducting market research is driven mainly by the fact that purchase intent can often not be surveyed directly from the end consumers of the technology (consider, for example, a new imaging method in cancer treatment, or a new insulation technology). Rather, market research is conducted via interactions with market experts, such as physicians in the healthcare industry, lead engineers for some industrial devices, or athletes for sports apparel. Identifying, locating and accessing these informants is costly in any market research effort, and particularly costly for early-stage technology teams who are often time and cash constrained and cannot explore each potential market in great depth, or even all of them at any depth. Our investigation reflects this feature by examining scenarios in which the number of alternatives is large, but the budget (i.e. the overall number of samples that can be collected) is small.

Classic bandit models use the sum of undiscounted rewards over a finite horizon, an average reward over the infinite horizon, or, more usually, a sum of discounted rewards up to the infinite horizon as the objective function (*Gittins et al.*, 2011). This is a suitable choice for many operational settings, in which the decision-maker must balance immediate profit generation and more risky, innovative activities that can increase future profits. In contrast, the "reward" for an early-stage team is the decision of an investor to fund them (or not) and the terms of that investment. In our model the signals collected by the team during their search do not have immediate payoff consequences. Rather, at the end of the market research phase the team selects the market that appears to be "best" and that market is used to generate their payoffs.

Because we are interested in search processes conducted by human decision-makers, our goal is to identify effective policies that can be translated into intuitive and easily implementable decision rules. This is different from most classic bandit investigations that focus on computational performance of policies (solution quality, speed, computational complexity), bypassing any implementability or intuitiveness aspects of the policy in practice. To evaluate implementability we will (a) recommend easy-to-communicate policies whenever their performance is sufficiently good, (b) examine the search process of less intuitive policies to distill the features that can be used to enhance simple policies and (c) stress-test each policy with regard to human errors revealing how robust the policies are to some characteristic behaviors of human decision-makers operating in stochastic environments. In particular, we will examine the robustness of policies to limited memory (*Gans et al.*, 2007), mental sampling (*Tong and Feiler*, 2016), misalignment of the prior beliefs and the true distribution of arms (overconfidence, *Herz et al.*, 2014), incorrect signal processing (also known as "overprecision", *Herz et al.*, 2014) and random errors ("tremble").[1]

To summarize, we will explore policy performance in three variants of the classic Bandit model revised to reflect key attributes of this entrepreneurial setting. These are: (a) each alternative exhibits a different risk profile, (b) the number of alternatives exceeds the number of samples that can be drawn, (c) the search performance is valued only by the selection made in the terminal stage, and (d) the search strategy needs be implementable by a human decision-maker.

To be able to inform entrepreneurial decision-making in this setting we will introduce three search models and examine the performance of a wide range of search strategies in a multitude of scenarios. In this dissertation chapter we will restrict our

[1]Overoptimism and overprecision refer to the decision-maker's incorrect perception of some or all arms being better (in terms of their means and sampling standard deviations) relative to the true information state. More broadly, these are manifestations of overconfident behaviors, which have been found to be particularly prevalent among entrepreneurs (*Busenitz and Barney*, 1997; *Forbes*, 2005; *Moore and Healy*, 2008; *Croson and Ren*, 2013).

analysis to scenarios defined within the context of the revised Bernoulli bandit model (model 1) and of the revised Normal bandit model (model 2). Most of the existing bandit policies were designed for either the Bernoulli or the Normal model, so testing policy performance in these standard models will reveal where some of the existing policies fail, once we account for the additional features of the entrepreneurial setting. In future research we plan to build on this analysis and examine policy performance in scenarios defined by the Risk-return model (model 3).

This is the first paper to our knowledge to formulate and study the entrepreneurial market identification problem. Our initial results offer some insights for technology-based startups seeking to commercialize their invention, and also add texture to some of the classic solutions to the exploration-exploitation dilemmas in the broader search and bandit literature.

First, we find that many classic index-based policies that perform well in conventional search settings fail in our setting. For example, Gittins-index based solutions, and Thompson sampling only work well with a small number of alternatives relative to the sampling budget. In contrast, another well-known class of strategies, *Stick-with-the-winner, Switch-from-a-loser* performs well across the different scenarios in our setting. While the implementation of these policies is trivial in search processes with binary market signals, it requires an adaptation step for non-binary signals. We explore a broad range of such adapted *Stick-switch* policies and find that policies with high threshold values for sticking with the current market (and a low threshold for switching to a new market) perform better than any other conventional search policy.

Second, we are able to further improve upon the performance of *Stick-switch* policies in settings with continuous signals, by adding a deep search stage at the end of the search horizon. This is particularly valuable in settings in which inherent market risk plays an important role. In these, a deep search stage can improve search performance by up to 20% relative to the stand-alone *Stick-switch* policy. The

relative advantage of such modified *Stick-switch* policies is confirmed in settings where the decision-maker is characterized by imperfect recall and information processing capacity.

The implication of these results in practice is that entrepreneurial teams should begin market research by spending a large portion of their budgets to the exploration of a broad range of potential markets, moving on quickly to new markets whenever they receive unfavorable information about a market. Towards the end of the market research phase teams should narrow down their search, deepening their understanding of the best market(s) discovered so far.

The remainder of this paper is organized as follows. Section 3.2 introduces the model. Section 3.3 describes the search policies. Section 3.4 presents the results of the simulations. Section 3.5 studies the implementation of the algorithms focusing on a boundedly rational decision-maker. Section 3.6 concludes.

3.2 Models

The iterative search and selection among multiple, risky alternatives has a mathematical expression in the literature: the multi-arm bandit model (*Gittins and Jones*, 1979; *Gittins*, 1979). In most existing applications of the bandit model, and in ours the decision maker decides dynamically on the actions to be taken while observing the outcomes of her past actions. In our setting, the decision-maker is the technology team. The arms of the bandit are the different markets or applications contained in $\mathbb{M} = \{1, 2, ..., M\}$. The team has a (time and/or cash) budget represented by N that determines how many samples they can collect to learn about the different markets. In stage $n = 0, 1, 2, ..., N-1$ the team chooses market $i_n \in \mathbb{M}$, observes signals $W_{i_n,n}$, which are correlated with the true arm performance (in ways that will be described later) and updates their beliefs about the desirability of arm i. In stage N the search is complete and the team chooses the arm that appears to be best ("best" will be

defined below).[2]

Since its conception, the bandit model has been extended in many directions to investigate exploration-versus exploitation trade-offs encountered in various operational settings. For example, *Bertsimas and Mersereau* (2007) consider a marketer learning the efficacy of alternative marketing messages, *Caro and Gallien* (2007) consider a retailer making assortment decisions and learning about demand, *Papanastasiou et al.* (2017) consider consumer populations learning (collectively) about the quality of service providers, and *Gans et al.* (2007) consider a manufacturer learning the quality of potential suppliers. However, to our knowledge none of the existing bandit models extrapolate directly to the entrepreneurial market research problem. We will next discuss the models in more detail.

3.2.1 Model 1 (Bernoulli Bandit)

The Bernoulli bandit model is one of the most studied Bandit variants, and will be used as our starting point for identifying well-performing search strategies. Each bandit arm (market) is uniquely defined by its probability of success, p_i. Then, the value of the technology in market i has two sources of variability or randomness: the inherent randomness in the fact that the number of sales in market i is a random variable, and the sampling error in the estimate of p_i. In our analysis we will ignore the inherent risk (because it is pre-ordained by parameter p_i and restrict our attention to the sampling error. Then, our objective is to find the market in which the technology will have the highest expected market share, while making sure that the market information is representative of the true market potential. Mathematically, the objective is to select iteratively markets $i^0, i^1, ..., i^{N-1}$ to identify the market that

[2]The entrepreneurial team may not be aware of some markets. One might then think of M as the number of markets the team was able to identify in their initial market research efforts.

exhibits the best performance in the terminal stage N:

$$\max_{i^0,i^1,...,i^{N-1}} v_i(p_{i^N}, \hat{p}_{i^N}^N, \lambda) = \lambda p_{i^N} - (1-\lambda)\sqrt{\text{Var}[\hat{p}_{i^N}^N]} \tag{3.1}$$

where $i^N = \text{argmax}_{i\in\mathbb{M}}\, \lambda p_i^N - (1-\lambda)\sqrt{\text{Var}[\hat{p}_i^N]}$.

3.2.2 Model 2 (Mean-variance Normal Bandit)

Model 1 begins to incorporate risk into policy performance evaluation, but the differences in market risk between the alternatives are restricted to the level of confidence (i.e. sampling variance) with which the decision-maker can predict market performance. To incorporate a richer characterization of risk, and to further ground our results in the existing Bandit literature we will also examine policy performance in a variant of the Normal bandit model with unknown mean and variance. In the Normal model each signal is used to update both the return estimate (mean of the distribution characterizing market i) and the risk estimate (variance of that distribution), which jointly determine the desirability of a market.

Let each market $i \in \mathbb{M}$ be characterized by a probability distribution $F_i \sim N(\theta_i, \sigma_i^2)$, where both the means θ_i and variances σ_i^2 are unknown. The objective of the team in this setting is to select iteratively markets $i^0, i^1, ..., i^{N-1}$ to identify the market that exhibits the best performance in the terminal stage N:

$$\max_{i^0,i^1,...,i^{N-1}} v_i(\theta_i, \sigma_i, \lambda) = \lambda\theta_{i^N} - (1-\lambda)\sigma_{i^N} \tag{3.2}$$

where $i^N = \text{argmax}_{i\in\mathbb{M}}\, \lambda\theta_i^N - (1-\lambda)\sigma_i^N$.

3.2.3 Model 3 (Risk-return Model)

While models 1 and 2 begin to account for some risk differences between the different markets, they do not fully reflect the richness of the interactions between the team

and their market informants. In model 3 we explicitly model the dual (risk+return) nature of the market signals obtained by the team during market research.

Let each market be characterized by two constants, S_i denoting the return that can be earned in market i, and R_i denoting the aggregate risk measure for market i.[3] The team initially knows little about S_i or R_i, other than that these quantities are non-negative and bounded. We will assume a non-informative prior for S_i on $[0, S^{max}]$, and similarly for R_i on $[0, R^{max}]$. We will denote the prior distributions by $f^S(s_i)$ and $f^R(r_i)$, respectively. If the team chooses market i, they receive two independent signals: one about S_i, and one about R_i. We will next describe the updating process for S_i; the updating for R_i follows the same logic and will be omitted for brevity.

Let $\tilde{S}_{i,n}$ be the market signals about S_i obtained in stage $n = \{1, 2, ..., N\}$ of the search process. We will assume that each signal $\tilde{S}_{i,n}$ is an unbiased but noisy estimate of the true quantity, S_i. We will denote the sampling noise by $\epsilon_{i,n}$ and assume it is stationary, denoting its pdf by $g^S(\epsilon_{i,n})$. We will further assume that the $\epsilon_{i,n}$ are bounded on $[-a, a]$, where $a < S^{max}/2$, and that $\mathbb{E}[\epsilon_{i,n}|S_i] = 0$. Further, the signals are additive in the true market size and the noise, $\tilde{S}_{i,n} = S_i + \epsilon_{i,n}$. Using Bayes' rule, the team can update their belief about market size S_i after observing signal $\tilde{S}_{i,n}$ as follows:

$$Pr(S_i = s_i|\tilde{S}_{i,n} = \tilde{s}_i) = f^S(s_i|\tilde{s}_i) = \frac{g^S(\tilde{s}_i - s_i|s_i)f^S(s_i)}{\int_t g^S(\tilde{s}_i - t|t)f^S(t)dt}, \tag{3.3}$$

with the updated belief $\mathbb{E}[S_i|\tilde{S}_{i,n}] = \int_{s_i} s_i f^S(s_i|\tilde{s}_i)ds_i$. In our baseline model we will assume that the prior density $f^S(s_i)$ and the noise distribution $g^S(\epsilon_{i,n})$ are uniform.[4]

[3] Another common criterion for venture evaluation is time-to-market (used for example, in the the Procter and Gamble three factor model of project evaluation, see *Cooper et al.*, 2006). In our model, time-to-market can be incorporated into the return measure by adjusting (delaying) the cash flow projections.

[4] For $S_i \in [0, a) \cup (S^{max} - a,\ S^{max}]$ the distribution of noise will be truncated at the bounds, with the consequence that the signal is not unbiased. In particular, we will assume that $g^S(\epsilon_{i,n}|s_i) \sim U[-a,\ a]$ for $S_i \in [a, S^{max} - a]$, that $g^S(\epsilon_{i,n}|s_i) \sim U[-s_i,\ a)$ for $S_i \in [0, a)$ and that $g^S(\epsilon_{i,n}|s_i) \sim$

The end point of the market assessment is the composite value of market i, v_i, which we define as the weighted sum of return and composite risk, $v_i = \lambda \times S_i - (1 - \lambda) \times R_i$, where λ is the weight put on the return measure relative to the risk measure. For the entrepreneurial team, the unknown quantities, S_i and R_i are replaced by their expectations, conditional on the knowledge state. The objective of the team is to select the measurements $i_1, i_2, .., i_{N-1}$ to discover the market with the highest value v_i. At the end of the market research phase, in stage N the entrepreneurial team selects the market with the highest $\mathbb{E}[v_i]$.

3.3 Search policies

As with many variations of this problem class exact solutions are elusive, but we will consider a range of heuristics that have been found to perform well in different contexts and test them in simulation. We will then qualitatively analyze the key structural components of well-performing heuristics and develop some new ones with the objective of being able to clearly communicate them to non-mathematical practitioners.

A key property of effective search strategies in traditional bandit models is the ability of the strategy to balance exploitation (earning rewards from arms known to perform well) and exploration (learning about the potential of new arms). Given that in our setting the signal realizations do not directly contribute to the earnings, one might expect the more exploratory policies to perform well. However, exploratory policies may suffer from high variance of the chosen arm. The interaction of these, and other model features for strategy performance is not obvious and has not been investigated extant bandit literature. We will study strategy performance in the simulations, using a variety of scenarios in which we vary the model parameters $(p_i, \theta_i, \sigma_i^2, N, M, \lambda)$.

$U[-a,\ S^{max} - s_i]$ for $S_i \in (S^{max} - a, S^{max}]$. The truncation at the bounds can reflect either the increased informativeness of extreme signals, or internal adjustments to overly pessimistic/optimistic signals falling outside of the bounds.

In addition we will examine policy performance when the decision-maker implementing the policy exhibits imperfect recall and updating of parameters when processing new information. Before discussing these scenarios in more detail we will survey the search policies used in the extant literature and discuss the modifications to these policies required by our setting.

The body of work on iterative search is now extensive, with most investigations using the bandit or ranking and selection frameworks to study computational performance of different policies (*Kim and Nelson*, 2006; *Gittins et al.*, 2011; *Powell and Ryzhov*, 2012). Most of these policies are well-functioning (in at least some context) heuristics, which have the advantage of being efficient to implement. Each policy π defines a (possibly stochastic) rule or function $V^{\pi}(S^n)$ mapping the state of the knowledge in stage n, S^n to the alternative(s) to be selected. The state of knowledge, S^n includes the parameters of the distribution describing the team's beliefs about each arm i in stage n. These parameters are $\{\hat{\alpha}_i^n, \hat{\beta}_i^n\}_{i=1,2,...,M}$ in the Bernoulli bandit scenario, and $\{\hat{\alpha}_i^n, \hat{\beta}_i^n, \hat{\theta}_i^n, \hat{\tau}_i^n\}_{i=1,2,...,M}$ in the Normal bandit scenario. In the Risk-return model the beliefs about the return and risk parameters will be denoted by $\{\hat{S}_i^n, \hat{R}_i^n\}_{i=1,2,...,M}$. For certain policies S^n may also be required to include the history of the previous draws, $\{i^j\}_{j=0,1,...,n-1}$.

Further, some policies will require as input the prior value estimates for each arm (i.e. the weighted sum of the mean and risk parameter estimates given the current information state). We will denote these quantities by $\hat{v}_i^n$. In the Bernoulli bandit scenario $\hat{v}_i^n = \lambda \hat{p}_i^n - (1-\lambda)\sqrt{\text{Var}[\hat{p}_i^n]}$. In the Normal bandit scenario $\hat{v}_i^n = \lambda \hat{\theta}_i^n - (1-\lambda)\hat{\sigma}_i^n$. In the Risk-return model $\hat{v}_i^n = \lambda \hat{S}_i^n - (1-\lambda)\hat{R}_i^n$.

3.3.1 Simple, time-invariant heuristics

We first discuss the set of simple, time-invariant heuristics with clear practical interpretation. Among these, perhaps the simplest policy is the *pure exploration*

strategy that samples each market with probability $1/M$ (This policy can also be adapted to sample randomly with replacement). The opposite of an exploration policy is the *pure exploitation* (Also known as greedy or myopic policy). This policy selects in each stage the arm with the highest expected performance (given the current state of knowledge), breaking ties randomly. This has the advantage of always acting to maximize the near term expected gain, but it may not explore enough to find the best arm.

Another intuitive policy is *Stick with a winner, switch from a loser*. We will sometimes refer to this policy as *Stick-Switch* or *SS*. This policy starts with a randomly selected arm, and continues pulling that arm until the first negative signal (failure) occurs. We will adapt the *SS* policy for continuous signal realizations (as in model 2) as follows: success (failure) will be defined as arm performance above (below) a certain threshold. The threshold can be a constant parameter determined ex ante, or an endogenous parameter determined dynamically. After the first failure a new arm is selected at random, and the process is repeated until the search budget is exhausted. If all arms have been sampled at least once, the arm with the highest success/failure ratio, or the arm with the fewest pulls is sampled (This scenario will not occur in our focal setting, where $M > N$). In the Risk-return model we will use the composite signal, $W_{i,n} = \lambda\tilde{S}_{i,n} - (1-\lambda)\tilde{R}_{i,n}$ to evaluate whether the signal is recorded as a success or as a failure. We will also explore a variant of the *SS* policy, *SS2*, which discards an arm after *two* consecutive failures, and a variant that uses the ratio of successes to failures to determine whether or not to switch to the new arm (which will be referred to as *Ratio (+/-)* policy).

3.3.2 Heuristics balancing exploration and exploitation

The simplest policy in this class divides the budget into an exploration phase in which each arm is sampled at random, and an exploitation phase in which current

high performers are sampled. This policy is sometimes referred to as the *Test-rollout* policy (*Schwartz et al.*, 2017). An alternative policy with a gradual transition from exploration to exploitation is the ϵ-*Greedy* policy. This policy mixes random exploration and exploitation in each stage, with pre-specified probabilities. It conducts random exploration with probability ϵ and chooses the best arm (using the greedy method) with probability $1-\epsilon$. The parameter ϵ is typically chosen to be a decreasing function of the search stage n (See *Sutton and Barto*, 1998, for further variations on this policy).

The next two policies are more sophisticated versions of the ϵ-*Greedy* policy. Rather than sampling at random, these policies attach performance-dependent weights to each arm. In particular, *Boltzmann* (sometimes referred to as *Soft max*) policy samples arm i with probability $q_i^{soft,n}$ proportional to its current predicted performance $\hat{v}_i^n$. The probability of choosing arm i is then given by

$$q_i^{soft,n} = \frac{e^{\rho \hat{v}_i^n}}{\sum_{i' \in \mathbb{M}} e^{\rho \hat{v}_{i'}^n}}, \tag{3.4}$$

where $\rho \in [0, \infty)$ is the "greediness" parameter (The policy becomes greedy as $\rho \to \infty$).

The *Randomized probability matching (Thompson sampling)* policy is similar to *Soft max*, except that the probability of choosing an arm is the probability of that arm being the best one among all arms (*Scott*, 2010). That is, we choose arm i with probability $q_i^{thom,n}$ given by

$$q_i^{thom,n} = \Pr\Big(\mu_i^n = \max\{\mu_1, \mu_2, ..., \mu_M\}\Big), \tag{3.5}$$

where μ_i is the mean of the distribution characterizing arm i (p_i in model 1, or θ_i in model 2). This method is typically implemented by drawing a sample from each arm's posterior distribution at random, and choosing the arm with the highest draw

value.[5]

3.3.3 Upper Confidence Bound (UCB) policies

The UCB class of policies was first introduced in (*Lai and Robbins*, 1985; *Lai*, 1987) who proposed to add an "uncertainty bonus" to the arm's empirical performance when computing that arm's index value. The size and the form of the uncertainty bonus may differ depending on the exact policy implementation (*Auer et al.*, 2002; *Audibert et al.*, 2008).[6]

Perhaps the most common variant of UCB is UCB1, given by:

$$v_i^{UCB1,n} = \hat{v}_i^n + \sqrt{\frac{2\log\, n}{N_i^n}}, \tag{3.6}$$

where N_i^n is the number of times arm i has been played up to and including time n. Notice that the second term is a decreasing function of the number of times an arm has been pulled.

Another frequently examined variant is UCB2. The main idea of UCB2 is to reduce the constant term in the fraction of time a suboptimal arm will be selected. In particular, UCB2 splits the search horizon into epochs of varying length. In each epoch the arm maximizing

$$v_i^{UCB2,n} = \hat{v}_i^n + \sqrt{\frac{(1+a)\log(e \times n/(1+a)^{r_i})}{(1+a)^{r_i}}}, \tag{3.7}$$

is selected and then played exactly $\lceil (1+a)^{r_i+1} - (1+a)^{r_i}) \rceil$ times before ending the epoch and selecting a new arm. The term r_i is a counter indicating how many epochs

[5]We also consider an adaptation of *Thompson* sampling, in which we use $\Pr\Big(v_i^n = \max\{v_1, v_2, ..., v_M\}\Big)$ instead of $\Pr\Big(\mu_i^n = \max\{\mu_1, \mu_2, ..., \mu_M\}\Big)$ to determine which arm should be played.

[6]UCB policies typically require each arm to be sampled at least once, or more than once before they can take action. Whenever possible we will adapt these policies to our setting by assuming that each arm has been sampled at least once or more times.

arm i has been selected in and $0 < a < 1$ is a tunable parameter.

UCB-tuned policy is similar to UCB1, but tunes the upper bound parameter in the padded term. The idea behind this policy is that the larger the total number of plays is, the more confident we should be before discarding a poor-performing arm. The value of an arm under this policy is given by

$$v_i^{UCB-tuned,n} = \hat{v}_i^n + \sqrt{\frac{2\log\ n}{N_i^n \min(1/4, \overline{Var}_i^n)}}, \tag{3.8}$$

where $\overline{Var}_i^n$ is an estimate of the upper bound of the variance of arm i.[7]

The KL-UCB approach is based on a padding function derived from the Kullback Leibler (KL) divergence (*Maillard et al.*, 2011).. KL-UCB considers the distance between the estimated distributions of each arm as a factor in the padding function. The KL-UCB policy operates by choosing the arm maximizing following expression:

$$v_i^{KL-UCB,n} = N_i^n d(\hat{p}_i^n, \mathbb{M}) \leq \log\ n + c\ \log\ \log\ n, \tag{3.9}$$

where $\mathbb{M}$ is the set of all possible distributions.[8]

Bayes-UCB (*Kaufmann et al.*, 2012) is the only UCB variant developed specifically for the Bayesian setting (the other UCB algorithms were developed for the frequentist setting, ignoring the prior distribution of the arms' values). The Bayes-UCB algorithm estimates quantiles of each arm's prior distribution to increasingly tight bounds. In each iteration Bayes-UCB draws the arm that maximizes the following expression:

[7] We use the Auer et al. 2002 approach, in which the estimate is given by $\overline{Var}_i^n = Var(\hat{v}_i^n) + \sqrt{\frac{2\log\ n}{N_i^n}}$.

[8] For Bernoulli arms $d(\hat{p}_i^n, \hat{p}_{i'}^n) = \hat{p}_i^n \log\frac{\hat{p}_i^n}{\hat{p}_{i'}^n} + (1 - \hat{p}_i^n)\log\frac{1-\hat{p}_i^n}{1-\hat{p}_{i'}^n}$ (*Kullback and Leibler*, 1951). Further, *Garivier and Cappe* (2011) find that setting the parameter $c = 0$ results in the best average performance. In our implementation we follow their recommendation.

$$v_i^{Bayes-UCB,n} = Q(1 - \frac{1}{n}, \hat{v}_i^n), \tag{3.10}$$

where $Q(t, \hat{v}_i^n)$ is the $t-$th quantile of the predictive distribution for the value of each arm.

The last policy in this category is DSEE (Deterministic Sequence of Exploration and Exploitation). This policy was introduced specifically for mean-variance bandits in *Vakili and Zhao* (2016). The idea is to explore at random in the first half of the planning horizon, and to use UCB1 in the second half. We also explore variations of this policy with other UCB-alternatives.

3.3.4 Forward-looking policies

The *m-step look ahead (Knowledge gradient, or POKER, or Price of Knowledge)* policy solves exactly the bandit problem if it were to end after m steps. The idea of collecting information based on the expected value of a single (or a few) measurements was first introduced by (*Gupta and Miescke*, 1994, 1996) and applied to the bandit setting in *Frazier* (2009). The m-step ahead policy chooses the arm i, with the highest value $v_i^{KG,n,m}$, defined as follows:

$$v_i^{KG,n,m} = \mathbb{E}\Big[\max_{i'} \hat{v}_{i'}^{n+m} - \max_{i'} \hat{v}_{i'}^n \mid \alpha_i^n, \beta_i^n\Big]. \tag{3.11}$$

Notice that the arm index in equation 3.11 maximizes the incremental improvement between stages n and $n+m$ and between the the best arm in stage n and in stage m, rather than the value of the arm that is expected to best in stage $n+m$. The reason is computational – the look ahead policies are computationally expensive, and the incremental improvement can often be approximated more easily than the predicted value of the best arm.(*Powell and Ryzhov*, 2012).

Several variants of the look-ahead policy that have been developed specifically for

the setting with unknown means and variances are "linear loss" (LL) policies. We will consider two of these policies: LL_1 and $LL(1)$. Both these policies have been shown to perform well in selection problems (*Chick and Inoue*, 2001; *Chick*, 2006; *Chick et al.*, 2010). Both policies use approximations to compute predictive distributions for the maximum improvement resulting from pulling each arm.

The m-step look ahead strategy with $m = N - n$ provides the optimal solution to the bandit problem. Solving this dynamic program is computationally intractable even for relatively small N. However, an optimal solution under some restrictive assumptions (Exponential family of distributions, infinite horizon, geometric discounting) is given by the *Finite-horizon Gittins index* Gittins index (*Gittins and Jones*, 1979; *Gittins*, 1979). While not necessarily optimal when these assumptions are not satisfied, Gittins indices have been shown to perform well in broader settings, for example in finite-horizon problems (*Niño-Mora*, 2011). In our setting the Gittins-index solution may not explore enough because our setting shifts all rewards to the terminal stage. Further, the solution must account for the role of risk. To compute Gittins indices for each arm we will use the dynamic programming solution concept of *Whittle* (1980).

3.4 Simulation results

3.4.1 Simulation setup

To evaluate policy performance we conduct simulations over a set values of the true parameters generated at random (in ways that will be described below). For each set of simulations we will conduct K runs of the simulation, and in each run $k = 1, 2, ..., K$ a new state of the world, or "truth" ψ_k is generated from the same distribution. In the Bernoulli bandit model the state of the world will be defined by the vector of true means for each market, $p^k = [p_1^k \; p_2^k \; ... \; p_M^k]$. In the Normal bandit

model the state of the world will be defined by the vector of tuples containing the true means and standard deviations: $\{\theta^k, \sigma^k\} = [\{\theta_1^k, \sigma_1^k\},\ \{\theta_2^k, \sigma_2^k\},\ \ldots\ , \{\theta_M^k, \sigma_M^k\}]$. In the Risk-return model the state of the world will be defined by $\{S^k, R^k\} = [\{S_1^k, R_1^k\},\ \{S_2^k, R_2^k\},\ \ldots\ , \{S_M^k, R_M^k\}]$. The matrix W^k collects the signal realizations $W_{i,n}^k$ for each alternative $i \in \mathbb{M}$ and all measurement stages $n = 1, 2, ..., N$. In each simulation run k we evaluate the value of the policy π by

$$F_k^\pi(\psi_k) = \max_{i \in \mathbb{M}} v_i(\psi_k) - v_{i^*}(\psi_k, \pi), \tag{3.12}$$

where $i^* = \text{argmax}_{i^N \in \mathbb{M}}\, \hat{v}_{i^N}^N$. Each sample path k is generated once and is used to evaluate each policy. The performance of a policy over K runs is then evaluated by:

$$\bar{F}^\pi = \frac{\sum_{\psi_k \in \Psi} F_k^\pi(\psi_k)}{K} \tag{3.13}$$

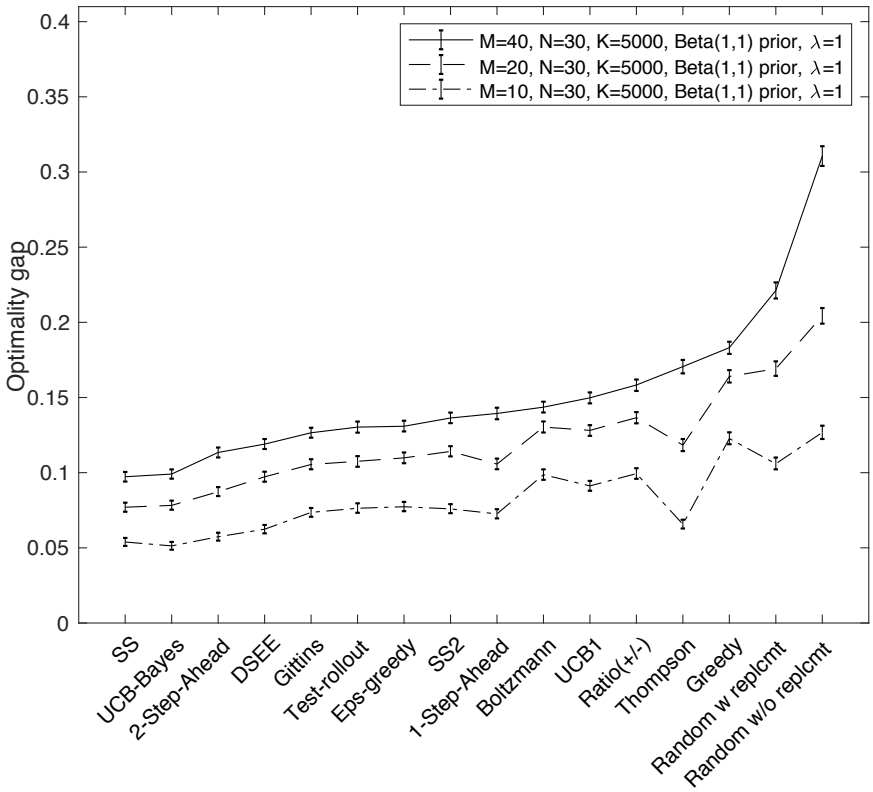
where Ψ is the set of true states of the world used in the simulations.

We will begin by examining the (conventional) bandit setting with few arms and many samples, and with the mean performance of the chosen arm being the sole performance metric. We will then examine the effects of two unique features of the entrepreneurial search problem on policy performance: the effect of the number of markets (relative to budget) and the effect of risk. We will then explore the search processes of well-performing policies and heuristics. The examination of policy performance under imperfect recall and updating is postponed until section 3.5.

3.4.2 Model 1: Informational risk (Bernoulli Bandit)

We first discuss simulation results for the Bernoulli bandit model (model 1). In the first comparison of the policies we examine the role of the number of arms (markets), M on policy performance. Figure 3.1 shows mean performance for each policy,

Figure 3.1: Policy performance, Bernoulli bandit



Note. Optimality gap (see equations 3.12 and 3.13) is used as performance metric. Policies are sorted by performance in the $M = 40$ scenario. Bars indicate bootstrapped 95% confidence intervals for mean performance. Heuristics that require parameter choices were parametrized iteratively in successive simulation runs: well-performing parameters are chosen.

along with the (bootstrapped) confidence intervals for mean performance.[9] Policy performance is evaluated using the optimality gap or regret in terms of the arm value (defined in equation 3.12). We vary the number of arms (markets) between 10, 20 and 40 in these simulations, keeping the number of samples, N constant, and keeping the weight on the mean estimate relative to the variance of that estimate constant ($\lambda = 1$). We use the uninformative Beta (1,1) prior distribution, which aligns with the true data generating process (misaligned priors will be examined in the next

[9]We created 5000 bootstrapped samples from actual performance data and used these samples to evaluate mean policy performance. The endpoints of the confidence intervals are the 2.5th and the 97.5th percentiles of that bootstrapped distribution. These intervals give a sense of statistical reliability of our simulation results.

section).[10]

Figure 3.1 reveals that the *Stick-Switch* and *UCB-Bayes* policies are the top performers across the different scenarios. Thompson sampling is relatively successful (on par with the top-performers) in the scenario with few arms, but performs very poorly in scenarios with many arms. Unsurprisingly, *Greedy* and *random* policies perform poorly across scenarios. Further, most of the policies that balance (in some way) exploration of new arms and exploitation of high performing arms perform better than *Greedy* or *random* policies, but worse than *Stick-Switch* or *UCB-Bayes*.[11]

Figure 3.2 shows mean performance for each policy keeping the number of markets constant, but varying the weights on mean and variance, λ. The *Stick-Switch* and *UCB-Bayes* policies remain the top performers with the mean-variance measure of policy performance. Among the remaining policies, *Thompson* policy exhibits the most noticeable deterioration in performance as the weights are shifted away from the mean and towards the variance. In contrast, *Ratio*, *Boltzman* and *Greedy* policies perform better, relative to other policies with lower values of λ.

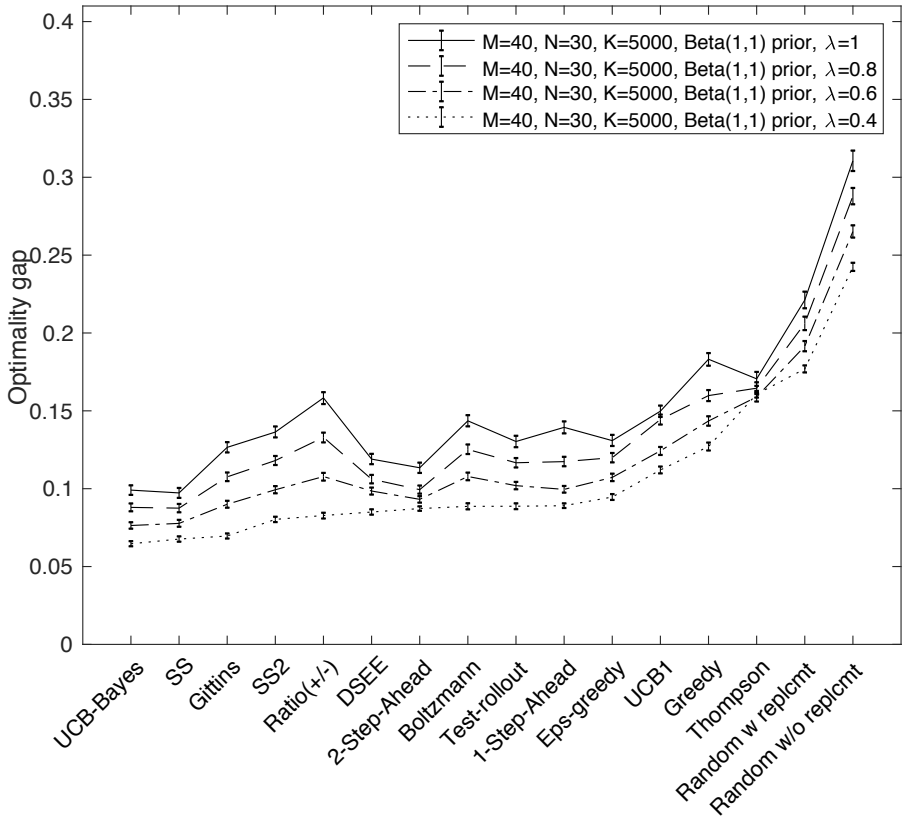
Search process

To examine the search process engaged by each strategy we will use the focal scenario with more arms than samples ($M = 40, N = 30$), and with a substantial weight on the variance relative to the mean ($\lambda = 0.6$). We examine policy behavior by looking at several variables related to the breadth and depth of exploration, and also at the changes in those measures over time. To examine the timing we look at policy behavior in the beginning (first half), vis-à-vis at the end (second half) of the search horizon. We focus in particular on the number of arms sampled, the number

[10] We repeat the analysis with negatively skewed priors (Beta(1,2); Beta (1,3) etc.) and find similar results. However, the differences between policies shrink for strongly skewed prior distributions, in our case, starting with Beta (1,3).

[11] These comparisons omit several *UCB* policies: *UCB2, KL-UCB, UCB-tuned*, whose performance resembles the performance of *UCB1* in most scenarios.

Figure 3.2: Policy performance, Bernoulli bandit



Note. Optimality gap (see equations 3.12 and 3.13) is used as performance metric. Policies are sorted by performance in the $\lambda = 0.4$ scenario. Bars indicate bootstrapped 95% confidence intervals for mean performance. Heuristics that require parameter choices were parametrized iteratively in successive simulation runs: well-performing parameters are chosen.

of arms sampled at least x times, the number of arms sampled in the first vs. second half, as well as on the number of reversals (a reversal is an event in which an arm is chosen that has been sampled previously but discarded in favor of another arm). We will next discuss how these quantities differ by search policy (the summary statistics are relegated to Appendix C).

Table C.1 (Appendix C) reveals that different policies can arrive at good performance results using somewhat different strategies. Consider the top-performing strategy, *UCB-Bayes*. Table C.1 suggests that it explores, on average 7 arms, chooses approximately half of those arms to conduct in-depth exploration (examining those arms at least twice), and does not change behavior between the first and the second

half of the search horizon. Interestingly, the second best search strategy, *SS* exhibits an almost identical search process as *UCB-Bayes*. Taken together, the results that both the performance and the search process of *UCB-Bayes* and *SS* are similar, as well as the intuitiveness of the *SS* policy suggest that it is a suitable candidate for communicating to human decision-makers.[12]

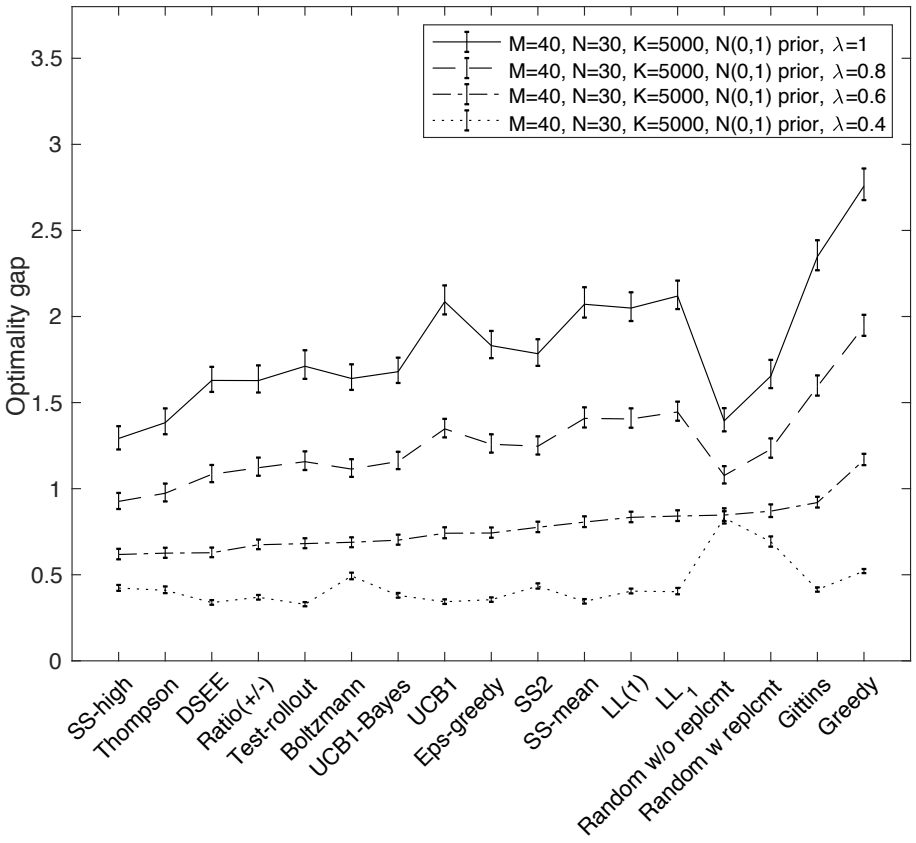
While *SS* and *UCB-Bayes* exhibit almost identical search processes, another successful policy, *Two-step-ahead* uses a very different approach: it exhibits frequent reversals, both in the first and the second half of the search horizon. In fact, we were not able to distill the search process of the *Two-step-ahead* policy to a human-implementable decision rule (We are not the first ones to find that step-ahead policies do not follow an easily distinguishable decision pattern (See, for example, *Powell and Ryzhov*, 2012, for a detailed discussion of interpretability of step-ahead policies).

One common feature of the three top-performing policies *(UCB-Bayes, SS, Gittins)* is that, despite the total number of arms being different between these policies, they explore approximately half of the sampled arms only once, and the other half at least twice). That is, when sampling variance is an important decision criterion, good policies explore approximately half of the arms superficially, and the other half in depth. Further, the exact number of the explored arms appears to be less important for policy performance than the ratio between the number of arms explored once and the number of arms explored at least twice. Stick-switch policies appear to be well-calibrated to achieve this balance, while other policies, such as UCB-Bayes and Gittins require more involved computation.

Lastly, the bottom 7 policies perform poorly because they either underexplore or over-explore, relative to the successful policies. Indeed, Table C.1 suggests that they sample either too few arms (*UCB1, Greedy, Ratio* policies), or too many (*Thomson,*

[12]The similarities between *UCB-Bayes* and *SS* are driven by the fact that a single negative signal results in the UCB-Bayes policy assigning a very low index to the attendant arm, causing that arm to never be used again.

Figure 3.3: Policy performance, Normal bandit



Note. Optimality gap (see equations 3.12 and 3.13) is used as performance metric. Policies are sorted by performance in the $\lambda = 0.6$ scenario. Bars indicate bootstrapped 95% confidence intervals for mean performance. Heuristics that require parameter choices were parametrized iteratively in successive simulation runs: well-performing parameters are chosen.

Random policies). In a setting, in which both the mean and the variance matter of the chosen arm matter, either too broad or too narrow exploration can lead to poor results.

3.4.3 Model 2: Inherent market risk (Normal Bandit)

We omit the discussion of the differences in policy performance for scenarios with varying number of arms (relative to samples): the relative ranking of the policies is not affected by the number of arms.[13] Instead, we focus on policy performance while

[13]One exception are Look-ahead policies: these perform well in Normal bandit settings with few arms, but not in settings with many arms.

varying the weights on the mean and the variance. Figure 3.3 shows the results for the $N(0,1)$ prior distribution for the unknown mean and variance parameters. In particular we examine scenarios with $\lambda = \{0.4, 0.6, 0.8, 1\}$. As before, we focus on the setting with more arms than samples ($M = 40$, $N = 30$).

Figure 3.3 reveals that *Stick-Switch* policies again, perform well relative to the other policies. In particular, *SS-high* is better than any of the remaining policies in scenarios with a moderate weight on the variance ($\lambda \in \{0.6, 0.8\}$). *SS-high* is a variant of the *SS* policy modified for the setting with continuous signals: it sticks with the current arm if the signal value is "sufficiently" high. *Stick-switch-high* uses the 95^{th} percentile of the prior distribution for the mean: any signal value above the 95^{th} percentile is interpreted as a success, and any value below that quantity is a failure. Our computational experiments suggest that such higher cutoffs perform better than lower cutoffs for moderate weights on the variance. Another *SS* variant, *SS-mean*, which uses the mean of the prior distribution of the mean to define success performs substantially worse than *SS-high*.[14]

Another policy, *Thompson sampling*, also performs well across the scenarios. It is worth mentioning that *Thompson sampling* did not perform well in the Bernoulli bandit scenario. One possible reason for the inconsistent performance of this policy is differences in the nature of the signals in the two models. With bounded signals *Thompson sampling* may undervalue strong signals, and may instead put too much probability mass on new, unexplored arms (Indeed, Table C.1 in Appendix C suggests that *Thompson* samples too many arms in the Bernoulli scenario). But, when signals are unbounded, *Thompson* sampling can allocate more probability mass to really good arms, and less probability mass to really bad arms, resulting in more nuanced decisions when choosing which arm to sample.

Lastly, it may seem surprising that random allocation with no replacement is a

[14] Additional experiments have shown that any value between the 90^{th} and 99^{th} percentile performs well in scenarios with ($\lambda \in \{0.6, 0.8\}$).

relatively effective strategy in the scenario with $\lambda = \{0.8, 1\}$. However, when variance plays a greater role in policy evaluation ($\lambda = \{0.4, 0.6\}$), random allocation of samples to arms is an ineffective strategy. The reason is that high signal realizations are likely to come from distributions with higher means, and are also likely to be generated by distributions with higher variances, which is penalized in scenarios with low λ-values.

Search process

Table C.2 in Appendix C summarizes the process variables for model 2 (Normal bandit) scenario. We again use the focal scenario with more arms than samples ($M = 40, N = 30$), and with a substantial weight on the variance relative to arm mean ($\lambda = 0.6$).

Let us first consider the number of arms sampled by each policy. Overall, successful policies search more broadly in the Normal bandit scenario, relative to the Bernoulli scenario. The top performing policy, *SS-high* exhibits a particularly broad search behavior, both in the first and in the second half of the search horizon. Further, most arms are sampled only once (as the threshold for sticking with an arm is set high). The result is that only 2.57 arms are sampled more than once, and only 1.08 arms are examined more than 3 times. That is, *SS-high* achieves good results by covering a substantial share of all available alternatives, and by narrowing down to one well-performing candidate arm.

Thompson sampling, the runner-up, exhibits a similar search process, though it searches somewhat less broadly relative to *SS-high*. Further, *Thompson sampling* examines fewer arms in the second half relative to the first half of the horizon and exhibits reversals, mainly in the second half of the horizon, suggesting that it follows a more conservative strategy in later stages of the search process. Another successful policy with a similar (although somewhat less broad) search process is DSEE.

The search process analysis suggests that good results can be achieved using very

different search strategies. Building on this insight we will next propose several new heuristics that combine some of the favorable properties of different policies.

3.4.4 Stick-switch based policies

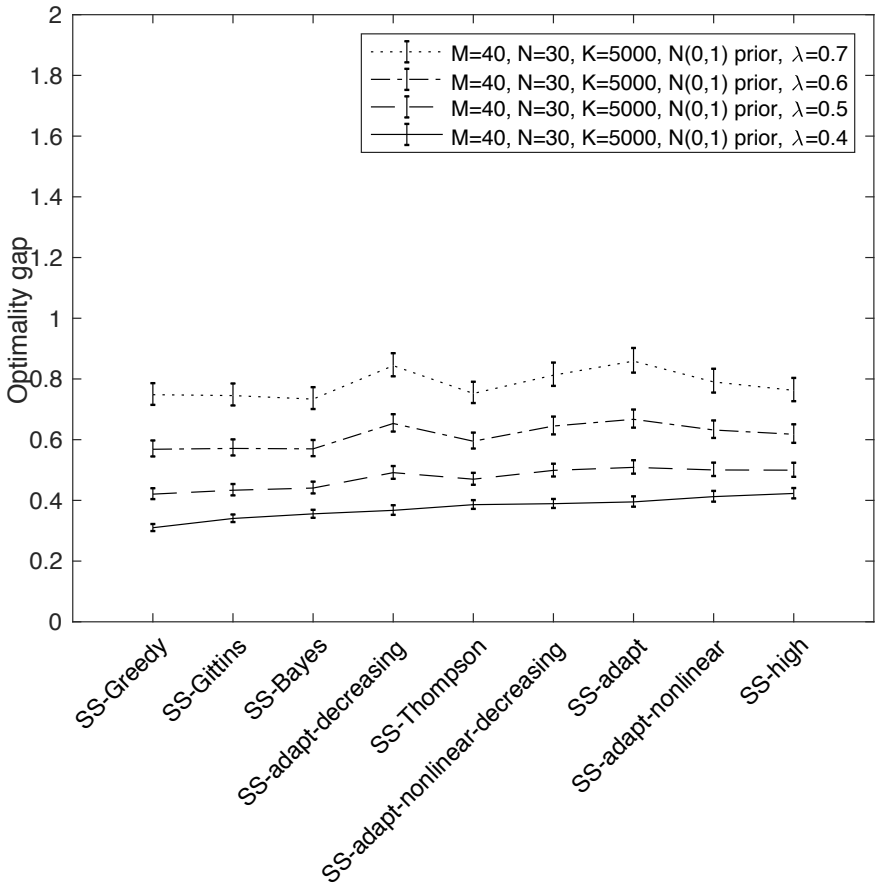
The search process analysis reveals that *Stick-switch* policies are time-invariant (in that they do not change the search strategy over the course of the search horizon), whereas some of the other high performing policies become less exploratory over time. We will use this insight to construct two types of modified *Stick-switch* policies: one that uses the *Stick-switch* approach in the beginning of the search, but switch to another, more conservative policy in the later stages of the search, and one that uses an adaptive threshold for deciding whether to stick with the current arm or to switch to a new arm.

In particular, we first construct policies that use the *Stick-switch* approach in the first N_1 stages of the search, where $1 \leq N_1 < N$ and another, more conservative policy, such as Greedy, or Gittins in the remaining $N - N_1$ stages.[15] These policies will be labeled *SS-Greedy*, *SS-Gittins*, etc. In addition, we examine adaptive *Stick-switch* policies, in which the switching threshold increases or decreases gradually over the course of the search horizon. These policies will be labeled *SS-adapt*.

Figure 3.4 summarizes the results for the Normal bandit model (In the Bernoulli bandit case the improvement relative to the stand-alone *Stick-switch* policy was insignificant). All of the modified *SS* policies perform better relative to the standalone *Stick-switch* policy. We find in particular, that *SS-Greedy* and *SS-Gittins* reduce the optimality gap by up to 15.8% relative to the standalone *Stick-switch* policy, and that the improvement is particularly big in the scenarios with a stronger weighting of the variance ($\lambda = 0.4$). Additional process analysis reveals that the improvement is

[15] N_1 can be determined deterministically, by examining alternative values of N_1, or it can be an endogenous parameter determined by the current arm value predictions. We use the former in the results presented in this section, and the latter as a robustness check. We use $N_1 = 25$ in the simulations discussed in this section, but find similar results for $N_1 \in [20, 27]$.

Figure 3.4: Stick-switch-based policies, Normal bandit



Note. Policy performance is measured as optimality gap (see equations 3.12 and 3.13). Policies are sorted by performance in the $\lambda = 0.5$ scenario. Bars indicate bootstrapped 95% confidence intervals for mean performance. *SS-Greedy, SS-Gittins* and *SS-Thompson* policies use the SS approach in the first 25 sample draws, and the other (*Greedy, Gittins* or Thompson) approach in the remaining 5 draws. SS-adapt denotes SS policies that increase the rejection threshold adaptively with the search stage n. In these, the change in the threshold can be being linear or nonlinear, as well as increasing (default) or decreasing in n.

achieved through reduced exploration and through a narrowing of the search to the few well-performing arms during the last few sampling stages.

Note in particular, that the most effective policies, *SS-Greedy* and *SS-Gittins* combine the broad exploration of the $Stick - switch$ approach in the first part of the search, and the exploitation approach in the later part of the search (*Greedy* and *Gittins* policies are the most conservative policies in the Normal bandit scenario, cf. table C.2 in Appendix C).

3.4.5 Discussion

Our investigation so far suggests that the *Stick-switch* policy and its derivatives are promising candidates to be shared with entrepreneurial teams conducting market research for a new technology.

In both the the Bernoulli and the Normal setting we saw that the performance of *Stick-switch* policies could be matched by more complex policies that involve exact computation and carrying forward of the prior and posterior distributions for each market's value. For example, in the Bernoulli model the *UCB-Bayes* policy performed on par with *Stick-switch* policies but required a computation of quantile functions with stage-dependent quantiles. However, we found that the search process engaged by *Stick-switch* policies was almost identical to *UCB-Bayes*, and that modified *SS* policies did not achieve substantial improvements over the standalone *Stick-Switch* approach.

While *Stick-switch* policies are easy to implement in settings with binary signals, the implementation for continuous signals is less clear. We have examined several *Stick-switch* policy variants adapted to the continuous signal setting and found that *Stick-switch* policies with high cutoff values (for classifying signals into "successes" and "failures") performed well across a variety of scenarios. Further, the performance of *Stick-switch* policies was stable for large deviations from the optimal parameters for the cutoff value. Overall, the robustness, and the intuitive nature of *Stick-switch* policies makes them desirable candidates to be communicated to market researchers for their implementation in practice.

In the Normal model *Thompson sampling* performed on par with the *Stick-switch* policy, but required a random device to select an arm in each stage, making it an unsuitable choice for sharing with a human decision-maker. However, we used one of the properties of the *Thompson sampling* strategy, frequent reversals to successful alternatives in the later stages of the search, to construct several new top-performing

policies. In particular, an additional performance improvement could be achieved by combining *Stick-switch* policies with conservative policies to help narrow down to a few promising candidate markets at the end of the search.

In addition to policy and parameter selection, another implementation concern is the correct execution of the algorithm by the decision-maker. We have so far considered a decision-maker whose priors align with reality, and who is able to update and remember all the prior and posterior distributions of the involved quantities when deciding which alternatives to sample. In the next section we will relax these assumptions.

3.5 Policy performance under imperfect recall and updating

3.5.1 Simulation setup

In this section we will examine the robustness of policies to imperfect implementation by the decision-maker. We will examine 4 types of cognitive limitations: limited memory, mental sampling, overconfidence (misaligned priors), as well as overprecision (placing too much weight on signals relative to the prior). These are some of the behaviors that have been considered in dynamic search processes, and more generally, in repeated stochastic decision-making environments (*Ederer and Manso*, 2013; *Herz et al.*, 2014; *Tong and Feiler*, 2016). We will also briefly discuss the role of random choices (sometimes referred to as "tremble" in the experimental economics literature *Bardsley and Moffatt*, 2007; *Moffatt*, 2016) in the search process. We will next describe how these behaviors are operationalized in the simulations.

"Limited memory" describes the scenario in which, rather than remembering the signals observed in periods $1, 2, ..., n-1$, the decision-maker in stage n only remembers the signals observed in periods $l, l+1, ..., n-1$, where $l > 1$. The posterior distribution of the parameters is constructed by updating the prior distribution for each arm with

the signals that are contained in the decision maker's memory.

"Mental sampling" describes the scenario in which, rather than using all the signals in her memory, the decision-maker uses a random sample of those signals to construct the posterior distributions for each arm from the prior distributions. If we denote the decision-maker's memory by L, and the size of her mental sample by O, where $O \leq L$, then the probability that any signal is used in the construction of the posterior distributions is O/L. The sampling is carried out without replacement, and occurs independently in each period.

"Overoptimism" describes the scenario in which the decision-maker begins the search with a more favorable prior distributions for some or all alternatives, relative to the truth. Denoting the degree of overoptimism about arm i by z_i, we construct overoptimistic prior distributions by endowing the decision maker with a $Beta(\hat{\alpha}_{0,i} + z_i,\ \hat{\beta}_{0,i})$ prior for arm i in the Bernoulli bandit scenario, and with a $\hat{\theta}_{0,i} + z_i$ prior for the mean parameter in the Normal bandit scenario. A decision-maker for whom $z_i > 0$ is overoptimistic.

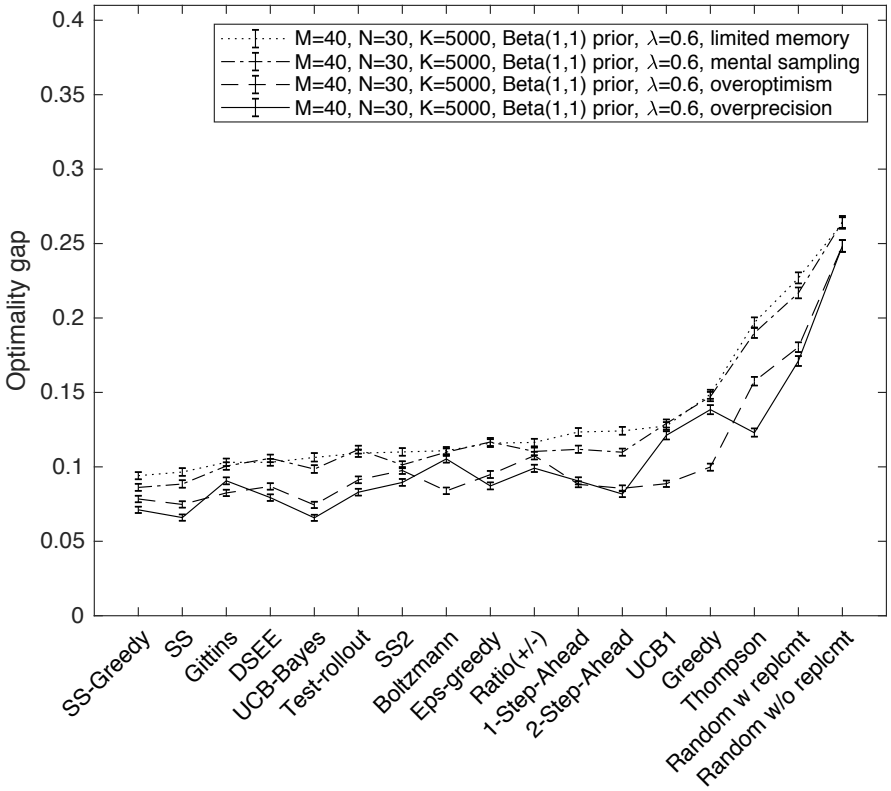
"Overprecision" describes the scenario in which the decision-maker places too much emphasis on the observed signals relative to the prior distribution. Note that unlike "overoptimism" which affects one's prior beliefs in stage 0, "overprecision" affects the updating at each stage of the search process. If we denote the degree of overoptimism by u_i, then the parameters of the $Beta-$distribution in the Bernoulli bandit model are updated as follows: $\hat{\alpha}_{n,i} = \hat{\alpha}_{n-1,i} + W_{n,i} \times u_i$, and $\hat{\beta}_{n,i} = \hat{\beta}_{n-1,i} + (1 - W_{n,i}) \times u_i$ in the Bernoulli model, and similarly for the Normal model (For the updating equations see, for example *Powell and Ryzhov*, 2012).[16]

3.5.2 Simulation results

We use the following parameters in the four scenarios discussed in this section:

[16] As a robustness check we also examine variations of this behavior, for example placing too much weight on positive signals but not on negative signals, and vice versa.

Figure 3.5: Policy performance under imperfect recall and updating, Bernoulli bandit



Note. Optimality gap (see equations 3.12 and 3.13) is used as performance metric. In the "limited memory" scenario, $L = N/2$. In the "mental sampling" scenario, $O = N/2$. In the "overoptimism" scenario $z_i = 1$ for $i \in \mathbb{M}$. In the "overprecision" scenario $u_i = 2$ for $i \in \mathbb{M}$. Policies are sorted by performance in the "limited memory" scenario. Bars indicate bootstrapped 95% confidence intervals for mean performance. Heuristics that require parameter choices were parametrized iteratively in successive simulation runs to choose well-performing parameters.

- "Limited memory": $L = N/2$, $O = N/2$, $z_i = 0$, $u_i = 1$ for all $i \in \mathbb{M}$
- "Mental sampling": $L = N$, $O = N/2$, $z_i = 0$, $u_i = 1$ for all $i \in \mathbb{M}$
- "Overoptimism": $L = N$, $O = N$, $z_i = 1$, $u_i = 1$ for all $i \in \mathbb{M}$
- "Overprecision": $L = N$, $O = N$, $z_i = 0$, $u_i = 2$ for all $i \in \mathbb{M}$

These parameters have been chosen to be consequential for policy performance without washing out all of the differences between policies. Further, in each scenario we assume that the recall and updating limitations affect both the search and the

selection stages of the process.[17]

Figure 3.5 suggests that most of our results hold in the Bernoulli bandit (model 1) case, even after when the decision-maker makes errors in the updating process. First, *Stick-switch* remains the best performing strategy in each scenario. This is not surprising given that the *Stick-switch* approach does not require the computation of posterior distribution, and only considers the signal realizations to decide which arm to choose next. Therefore, when the *Stick-switch* policy is employed, the recall and updating errors only affect the performance in the last round, in which the policy selects the best arm to be chosen for the reward.

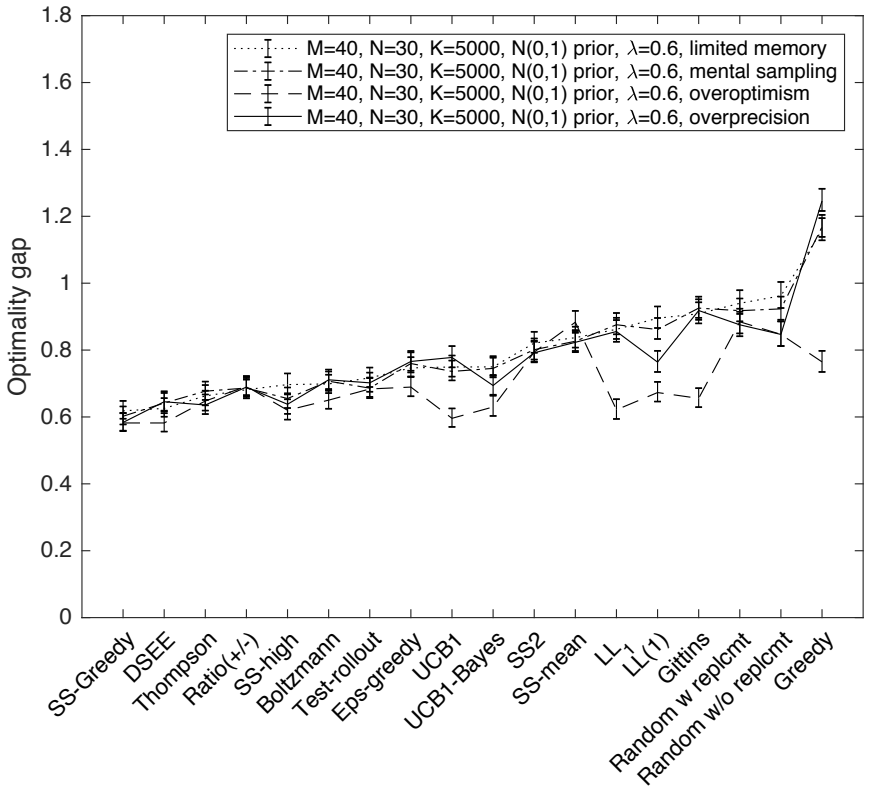
In contrast to the *Stick-switch* approach, the *UCB-Bayes* policy is affected in the "limited memory" and in the "mental sampling" scenarios (but not in the "overoptimism" and "overprecision" scenarios). That is, correct computation of quantile functions required by the *UCB-Bayes* policy appears to be sensitive to wrong parameter values. Another computationally expensive strategy, *Gittins*-policy performs well under a variety of assumptions, except for the scenario in which the decision-maker places too much emphasis on recent signals ("overprecision").[18]

Figure 3.6 shows the effects of imperfect recall and updating in the Normal bandit model. Again, the relative ranking of the policies is almost unchanged in the "limited memory" and "mental sampling" scenarios. In particular, *SS-Greedy*, which was found to outperform the remaining policies in the perfect updating and recall scenarios continues to be a top performer. Interestingly, some policies' performance improves under overoptimism (In particular, *UCB* and *Gittins* policies), relative to their performance in the baseline scenarios. This suggests that these policies may be adapted to perform better in our setting by changing the selection mechanisms they

[17]An alternative would be to assume that the decision-maker suffers from imperfect recall and updating *during* the search but is able to look back at the entire sequence of signals at the end of the search. We examine this possibility in a robustness check and do not find qualitative differences from the results presented in this section.

[18]We conduct additional simulations, in which only a subset of all arms was affected by "overoptimism" and found similar results.

Figure 3.6: Policy performance under imperfect recall and updating, Normal bandit



Note. Optimality gap (see equations 3.12 and 3.13) is used as performance metric. In the "limited memory" scenario, $L = N/2$. In the "mental sampling" scenario, $O = N/2$. In the "overoptimism" scenario $z_i = 1$ for $i \in \mathbb{M}$. In the "overprecision" scenario $u_i = 2$ for $i \in \mathbb{M}$. Policies are sorted by performance in the "limited memory" scenario. Bars indicate bootstrapped 95% confidence intervals for mean performance. Heuristics that require parameter choices were parametrized iteratively in successive simulation runs to choose well-performing parameters.

employ.

In addition to the results presented in figures 3.5 and 3.6 we conducted simulations in which the decision-maker was assumed to implement the policy imperfectly, choosing an arm at random in some rounds, instead of choosing the arm according to the algorithm (introducing "tremble" into the selection process). These simulations showed a general reduction of the differences in policy performance, but did not affect the relative ranking.

3.5.3 Discussion

The investigation of policy performance under imperfect recall and updating confirms the advantage of *Stick-switch* policies over the remaining policies. Because *Stick-switch* policies do not require iterative updating of posterior distributions of the arms, they survive some of the cognitive limitations that can be expected when the search is conducted by humans. In contrast, many other policies are negatively affected by these types of recall and updating errors, resulting in the deterioration of their performance. For example, some of the more computationally intensive policies, such as *Gittins* or *UCB-Bayes*, exhibit reduced performance. The robustness of *Stick-switch* policies to human implementation errors suggests that they are good candidates for being shared with decision-makers in practice.

3.6 Concluding remarks

This is one of the first papers to formulate and examine the entrepreneurial market identification problem. We first revised the bandit model to reflect the key features of the entrepreneurial setting. We then used a novel approach to identify solutions to this problem, which involved simulation and interpretation of successful search strategies, with the goal of sharing these strategies with entrepreneurial decision-makers in practice.

We first evaluated policy performance in the Bernoulli bandit setting and found that the intuitive *Stick with the winner - switch from a loser* strategy performed at least as well or better than any other conventional search policy.

We then adapted the *Stick-switch* logic to the setting with continuous signals and found that the resulting search algorithm, again, outperformed all of the remaining policies. Examining the search process engaged by different policies, we were able to further improve performance of the *Stick-switch* approach. This was achieved by

adding a deep search stage at the end of the search horizon, during which the best discovered alternatives are sampled repeatedly. The resulting approach is effective in a multitude of settings, robust to deviations from the assumptions of perfect recall and updating, and easy to communicate to decision-makers in practice.

Our results are intriguing given the emphasis of the search literature on index policies designed for optimal or near-optimal computational performance and not for human implementation. We find that in our setting such index-based policies do not outperform simpler heuristics, such as the *Stick-switch* algorithm. One of the drivers of this result is the fact that most index policies were designed to deal with a setting with few alternatives and many samples. But, when the number of alternatives to consider is a decision variable itself (as is the case when hypotheses outnumber samples), these policies fail to capture the relevant trade-offs.

The next steps for this investigation are twofold. First, the effectiveness of *Stick-switch* policies needs to be validated with humans. This will involve first communicating the policy mechanism to non-mathematical decision-makers and then examining its implementation and performance relative to some other (potentially decision-makers' own) search strategies.

Second, our computational results focus on settings with (ex ante) identical and uncorrelated alternatives. However, in some settings a different model of the underlying market landscape may be more appropriate. For example, markets may be divided into clusters, with strong within-cluster similarities. In this setting a signal from one market in a cluster may reveal something about some other markets. We also do not examine the setting in which both the inherent risk and the sampling uncertainty affect policy evaluation. Whether or not relative policy performance will change in these scenarios is open.

APPENDICES

APPENDIX A

Additional Tables for Chapter I

A.1 Participant demographics

Table A.1: Demographic variables (treatment means)

Treatment	College major				Age	Gender (1=f)	Performance ($)
	Social sci, arts, humanities	Bus, law, econ	Sci, med	Engineering, architecture			
Endog	0.36	0.14	0.41	0.09	23.27	0.64	3.39
5/15	0.28	0.13	0.52	0.07	21.52	0.39	5.17
10/10	0.48	0.06	0.23	0.23	22.52	0.56	6.28
15/5	0.41	0.03	0.41	0.14	22.07	0.52	5.38
Nudge	0.25	0.28	0.30	0.17	20.57	0.65	5.53
Pre-commit	0.16	0.27	0.25	0.28	22.63	0.63	6.07
Prototype	0.25	0.17	0.38	0.20	21.84	0.31	6.67
Total	0.31	0.16	0.35	0.17	22.09	0.53	5.49

A.2 Exogenous and Endogenous treatments, process variables

Table A.2: Design activity variables: summary statistics

	Treatment means				Treatment means		
	5/15	10/10	15/5	$p-$value	Exog	Endog	$p-$value
Count Variables							
# ideas	1.79	1.89	2.09	0.229	1.94	1.63	0.067
# elements	3.89	3.82	4.09	0.312	3.94	3.69	0.318
# all collapses	2.56	2.90	2.46	0.982	2.63	1.80	0.033
# coin collapses	1.65	2.05	1.82	0.865	1.85	1.18	0.029
# other collapses	0.90	0.86	0.65	0.930	0.79	0.61	0.248
# all coin stackings	5.65	5.51	5.95	0.895	5.71	4.38	0.016
# successful stackings	4.00	3.46	4.13	0.738	3.87	3.19	0.140
Time variables							
Time-to-first idea	02:51	04:36	02:55	0.704	3:28	4:39	0.017
Time-to-first collapse	06:32	08:06	07:25	0.477	7:26	9:24	0.091
Time-to-first stacking	05:44	06:52	04:27	0.627	5:38	8:18	0.011
Time-to-last idea	06:22	07:43	08:05	0.156	7:28	8:05	0.721
Time-to-last collapse	13:00	14:20	13:46	0.836	13:47	13:51	0.924
Time-to-last stacking	16:28	16:57	18:29	0.009	17:23	16:37	0.415

Note. Columns 2-4 and 6-7 show means of activity variables by treatment. Reported $p-$values indicate significance levels from Trend tests for 5/15, 10/10, 15/5 comparisons and two-sided Rank Sum tests for Exog vs Endog comparisons.

A.3 Multiple Hypothesis Adjustment

Table A.3: Multiple hypothesis adjustment

Analysis	Variable	Coef	Unadjusted $p-$value	Adj. $p-$value (*Holm*, 1979)
Endog and Exog treatments:	5/15	1.416	0.009	0.027
Treatment effect on	10/10	0.794	0.046	0.092
non-failure (Table 2, col. 2)	15/5	0.395	0.265	0.265
Endog and Exog treatments:	5/15	3.688	0.034	0.068
Treatment effect on	10/10	4.181	0.011	0.033
performance (Table 2, col. 6)	15/5	3.093	0.056	0.068
Endog, Exog and	Exog	3.584	0.010	0.040
additional treatments:	Nudge	2.937	0.082	0.156
Treatment effect on	Pre-commit	2.908	0.078	0.156
performance (Table 4, col. 1)	Prototype	4.367	0.010	0.040
Endog, Exog and	Exog	3.119	0.022	0.066
additional treatments:	Nudge	3.120	0.057	0.104
Joint effects of treatments	Pre-commit	3.112	0.052	0.104
and process variables	Prototype	4.951	0.003	0.012
(Table 4, col. 2)	Time-to-first idea	-0.490	0.000	0.000
Endog, Exog and	Exog	2.498	0.069	0.207
additional treatments:	Nudge	2.241	0.172	0.258
Joint effects of treatments	Pre-commit	2.429	0.129	0.258
and process variables	Prototype	3.856	0.020	0.080
(Table 4, col. 3)	Time-to-first stacking	-0.416	0.000	0.000
Endog, Exog and	Exog	4.264	0.008	0.024
additional treatments:	Nudge	2.871	0.152	0.152
Joint effects of treatments	Pre-commit	4.103	0.030	0.060
and process variables	Prototype	6.587	0.001	0.004
(Table 4, col. 4)	Time-to-first collapse	-0.400	0.000	0.000

Note. The adjusted p-values are calculated for each "family" of hypotheses. We draw on the definition of the family of hypotheses in *List et al.* (2016). We define the "family" of hypotheses as the group of tests of the effects of multiple treatments (and additional covariates in question) on the same outcome variable, in our case binary or continuous measures of performance. We use the Holm-Bonferroni adjustment (*Holm*, 1979). This procedure is a sequential version of the Bonferroni correction. We first obtain the unadjusted $p-$values. The hypotheses are then ordered from the one with the smallest p-value to the one with the largest. The hypothesis with the lowest p-value is tested first using the standard Bonferroni correction. The second $p-$value is then adjusted using the Bonferroni correction but the number of hypotheses is reduced by one. The same procedure is repeated for the remaining $p-$values.

A.4 Instructions [Exact Transcript, Endog Treatment]

Your objective is to build a structure that will support as many coins as possible as high off the table as possible. Your structure may use at most 10 cards and 10 clips. You will have a total of **20 minutes** to complete this task. Please raise your hand if you finish working earlier, so that the experimenter can evaluate your work.

Your Payoff.

Your performance will be judged based on the following formula:

$$\text{Your Payoff} = \frac{[\text{height of the highest set of coins in inches}] \times [\text{monetary value of these coins}]}{3}$$

The coins that count toward your payoff include the highest stack of coin (measured as the distance between the highest coin and the table), and all other coins at the same height level as this stack. The height will be rounded to the nearest inch. For example, if your highest coin is 9 inches off the table and there is a total of 8 coins stacked at that height, you will receive $\frac{9\times8\times\$0.25}{3} = \$6$ for this task. Please keep in mind that your structure has to be stable, so that the experimenter can measure the height reliably. To be precise, your structure has to stand for at least 3 minutes. If it collapses within 3 minutes after submission, your payoff for this task will be 0.

Note that if you place coins at different heights, coins that are not at the same level as the highest set of coins will not count towards your payoff. For example if your structure is 9 inches high, but you have placed 8 quarters at the top and 5 quarters at the height of 2 inches, your payoff will only include the value of the 8 quarters at the top. Thus, your payoff will still be $\frac{9\times8\times\$0.25}{3} = \$6$. In other words, only the set of coins at the highest distance off the table counts. You are not allowed to distribute the coins over multiple structures.

Timing.

Completion of the task consists of two parts: Design and Implementation. For the design part you will get an unlimited amount of playing cards and clips. The design materials should help you explore different possibilities. Experimenting may improve the final outcome of your work. Make sure that you make the most out of the materials you are given.

Once you feel certain about the final structure you want to submit, raise your hand. The experimenter will then take away your first set of materials and give you the final set of materials. Now the set of materials will include 10 cards and 10 clips only. These are the materials that you will use for the implementation.

You will have a total of 20 minutes, which means you must plan ahead, so that you have enough time to build your final structure. For example, if you raise your hand after 10 minutes, you will have 10 minutes left to implement your design using the final set of materials. It is your responsibility to tell the experimenter when you want to get the final set of materials, so that you can build your final structure.

APPENDIX B

Additional Tables for Chapter II

B.1 Estimation results for negotiation process in DELAYED treatment.

This appendix presents detailed estimation results discussed in Chapter 2.5. Table B.1 shows coefficient estimates for the regression of the probability of a contract offer being accepted (vs. being rejected or receiving a counteroffer) on the proposer's stage I contribution. Predictive margins in figure 3 are computed using this specification.

B.2 Subject heterogeneity by type.

The following tables present detailed estimation results discussed in section 2.6. Table B.2 shows coefficient estimates for the regression of the probability of a subject expressing a preference for a certain contract form and of a subject selecting a certain contract on the subject's type. Table B.3 shows coefficient estimates for the regression of effort on both contracts and subject type. Table B.4 shows within-type effort changes between contracts and type-specific response to partner effort. In all three

Table B.1: Effects of proposer's stage I contribution on contract acceptance decisions

Dependent Var.: *Response*		Sample: EQUAL offers	Sample: THRESH offers	Sample: DIFF offers	Sample: PROP offers
Accept contract (Response = 0)		(baseline)	(baseline)	(baseline)	(baseline)
Offer another contract (Response = 1)	*Proposer's st.I contribution*	-0.170**	-0.032	0.077	0.007
		(0.083)	(0.052)	(0.060)	(0.046)
	Constant	0.208	-1.992	0.851	2.102*
		(1.586)	(2.089)	(1.509)	(1.273)
Reject contract (Response = 2)	*Proposer's st. I contribution*	-0.097*	-0.030	0.023	-0.020
		(0.054)	(0.050)	(0.044)	(0.037)
	Constant	1.212	1.281	0.303	0.723
		(0.969)	(1.796)	(1.024)	(0.693)
	Observations	87	125	138	138

Note. Dependent variable is the response to a contract offer (0: Accept, 1: Offer a different contract, 2: Reject). Estimation is conducted using Multinomial Logit model, standard errors clustered at subject level. Each column uses observations in which a given contract type was offered. Coefficients are reported in relative risk ratio format. Age, gender and period are controlled for.

tables estimation is conducted using experimental data from experimental rounds 4-8 (to separate the analysis of type behaviors from type assignment for which rounds 1-3 data was used).

For robustness the analyses presented in tables B.2- B.4 have been replicated using experimental data from rounds 1-8. In a further robustness check we regressed effort on type and contract variables and all pair-wise interactions between contracts and types. For further robustness, we repeat the analysis interacting the type variable with the incentive strength of the contract (taking the value 1 for EQUAL, 2 for VESTING and 3 for PROPORTION contract). These robustness checks confirm our

results.

Table B.2: Contract preferences and contract choices, by type.

Contract		Dependent variable: Contract preference	Dependent variable: Final contract
EQUAL	*Conditional contributor*	-3.106***	-0.804
		(0.904)	(0.630)
	High contributor	-16.272***	-15.010***
		(0.905)	(0.718)
	Constant	-0.538	-3.561**
		(1.279)	(1.413)
VESTING		(baseline)	(baseline)
PROPORTION	*Conditional contributor*	-1.953**	-1.155*
		(0.835)	(0.662)
	High contributor	0.084	0.607
		(0.909)	(0.772)
	Constant	-2.004	-0.830
	Subjects	54	54
	Observations	270	270
		Tests of lin. com. of coefficients	
EQUAL	*High contributor − conditional contributor*	NA	NA
PROPORTION	*High contributor − conditional contributor*	7.664***	5.825***
		(4.477)	(3.317)

Note. Dependent variable is the expressed preference for a contract form (column 1) and final contract selected by the team (column 2). Preference for a contract is measured as the first offer made in negotiations. Estimation is conducted using Multinomial Logit model, standard errors clustered at subject level. Coefficients are reported in relative risk ratio format (i.e. the ratio of type-specific choice probabilities for different contracts). $VESTING$ contracts and Low contributors are used as the baseline. All coefficients are estimated using data from experimental rounds 4-8 (to separate type identification conducted in rounds 1-3 from type behavior). The bottom panel of the table shows tests of linear combinations of coefficients (again using the risk ratio format). NA denotes tests for which there is an insufficient number of observations of a type in a certain contract. Age, gender and experimental period are controlled for.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.3: Effort, by contract and type.

	Dependent variable		
	Stage I effort	Stage II effort	Total effort
VESTING	22.467***	32.962***	27.316***
	(7.354)	(8.860)	(6.885)
PROPORTION	30.036***	42.412***	35.622***
	(7.811)	(9.005)	(7.491)
Conditional contributor	14.748	18.162**	16.424*
	(9.174)	(8.235)	(8.602)
High contributor	18.099*	20.447**	19.365**
	(10.012)	(9.620)	(9.758)
Constant	16.563	-0.448	8.532
	(15.345)	(14.848)	(13.875)
Observations	270	270	270
Subjects	54	54	54
	Tests of linear combinations of coefficients		
VESTING −	-7.569**	-9.451***	-8.306***
PROPORTION	(3.651)	(3.630)	(2.895)
High contributor −	3.35	2.285	2.941
Conditional contributor	(6.504)	(7.063)	(6.725)

Note. Dependent variable is effort (stage I effort in column 1, stage II effort in column 2, total effort in column 3). Baseline is low contributor and EQUAL contract. Estimation is conducted using random effects regression model. All coefficients are estimated using data from experimental rounds 4-8 (to separate type identification conducted in rounds 1-3 from type behavior). Standard errors are clustered at subject level. Age, gender and experimental period are controlled for.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table B.4: Within-type, between contract effort comparisons.

	Dependent variable: Stage II effort					
Sample:	Low type	Conditional type	High type	Low type	Conditional type	High type
EQUAL	(baseline)	(baseline)	NA	(baseline)	(baseline)	NA
VESTING	29.993**	32.253***	(baseline)	30.130**	30.229***	(baseline)
	(13.831)	(11.219)		(13.532)	(11.419)	
PROPORTION	42.362***	45.027***	2.269	41.045**	41.855***	1.843
	(15.719)	(11.394)	(3.325)	(16.414)	(11.895)	(3.121)
Partner stage I contribution				0.186	0.629***	0.328
				(0.320)	(0.212)	(0.262)
Constant	4.787	28.755*	24.641	2.300	28.970*	23.323
	(103.814)	(14.900)	(26.221)	(106.659)	(15.168)	(25.390)
Observations	45	175	50	45	175	50
Subjects	9	35	10	9	35	10
Tests of linear combinations of coefficients						
VESTING −	-12.370	-12.770***	NA	-10.910	-11.630**	NA
PROPORTION	(13.860)	(4.578)	NA	(13.630)	(4.642)	NA

Note. Dependent variable is stage II effort. Baseline for low and conditional contributors is EQUAL contract. Baseline for high contributors is VESTING contract (there is not a sufficient number of observations of high contributors in EQUAL contracts). Estimation is conducted using random effects regression model. All coefficients are estimated using data from experimental rounds 4-8 (to separate type behaviors from type identification conducted in rounds 1-3). NA denotes tests for which there is an insufficient number of observations of a type in a certain contract. Standard errors are clustered at subject level. Age, gender and experimental period are controlled for.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

APPENDIX C

Additional Tables for Chapter III

C.1 Process analysis for Bernoulli and Normal bandit models.

Table C.1: Bernoulli Bandit: Search process for $Beta(1,1)$ prior distribution, $\lambda = 0.6, N = 30, M = 40, K = 10000$

Policy	Optim. gap	Arms sampled	Arms sampled 2+ times	Arms sampled 3+ times	Arms sampled in 1st half	Arms sampled in 2nd half	Reversals	Reversals in 1st half	Reversals in 2nd half
UCB-Bayes	0.076	7.26	3.56	2.32	4.75	3.32	0.00	0.00	0.00
SS	0.078	9.04	4.42	2.89	5.49	4.30	0.00	0.00	0.00
Gittins	0.090	4.88	2.44	1.62	4.24	1.56	0.00	0.00	0.00
KG2	0.093	10.76	5.06	3.18	6.89	5.84	12.60	6.17	6.43
DSEE	0.099	15.00	2.41	1.60	15.00	2.42	2.45	0.00	2.45
SS2	0.099	4.30	2.62	2.25	3.10	2.10	0.00	0.00	0.00
KG2	0.100	13.06	5.24	3.06	7.10	8.26	12.81	6.24	6.58
Test-Rollout	0.102	12.65	3.42	1.69	12.64	1.94	3.93	2.00	1.93
Eps-Greedy	0.107	12.66	3.72	1.88	8.26	6.52	8.76	3.67	5.09
Ratio (+/-)	0.108	3.78	1.89	1.25	3.46	1.28	0.00	0.00	0.00
Boltzmann	0.108	3.80	1.87	1.41	3.49	1.37	0.81	0.47	0.34
UCB1	0.124	3.02	1.50	1.50	2.76	1.23	0.00	0.00	0.00
Greedy	0.143	2.71	1.35	1.35	2.60	1.10	0.00	0.00	0.00
Thompson	0.159	22.14	5.73	1.52	12.98	12.49	7.12	1.70	5.41
Random w rpl	0.192	21.32	6.86	1.53	12.64	12.65	7.97	2.00	5.96
Random w/o rpl	0.265	30.00	0.00	0.00	15.00	15.00	0.00	0.00	0.00

Note. Policies are sorted by their performance (Optimality gap, column 2). Optimality gap denotes the distance between the best arm (in terms of value v_x) and chosen arm. Reversals are samples drawn from an arm that has been previously played and discarded in favor of another arm.

Table C.2: Normal Bandit: Search process for $N(0,1)$ prior distribution, $\lambda = 0.6, N = 30, M = 40, K = 10000$

Policy	Optim. gap	Arms sampled	Arms sampled 2+ times	Arms sampled 3+ times	Arms sampled in 1st half	Arms sampled in 2nd half	Reversals	Reversals in 1st half	Reversals in 2nd half
SS-high	0.618	23.75	2.58	1.08	12.29	11.70	0.00	0.00	0.00
Thompson	0.625	19.48	3.57	1.37	12.11	9.33	5.41	1.49	3.92
DSEE	0.628	15.89	3.27	1.96	15.00	3.89	4.08	0.00	4.08
Ratio (+/-)	0.674	13.46	2.41	1.13	8.97	5.10	0.00	0.00	0.00
Test-Rollout	0.681	15.89	4.82	1.85	12.65	5.94	5.55	2.00	3.55
Boltzmann	0.689	16.46	5.78	2.42	10.64	9.19	9.01	2.77	6.23
UCB-Bayes	0.701	11.49	2.73	1.81	6.72	5.41	0.01	0.00	0.00
UCB1	0.741	9.48	3.63	2.30	5.78	4.42	0.00	0.00	0.00
Eps-greedu	0.742	12.54	4.08	2.14	8.13	6.53	8.50	3.50	5.00
SS2	0.776	12.26	11.97	1.32	6.70	6.35	0.00	0.00	0.00
SS-mean	0.806	9.03	4.42	2.88	5.49	4.29	0.00	0.00	0.00
$LL(1)$	0.833	8.87	2.50	2.00	4.72	5.13	1.63	0.32	1.30
LL_1	0.841	10.01	4.00	2.97	6.51	5.18	15.27	6.63	8.64
Random w/o rpl	0.847	30.00	0.00	0.00	15.00	15.00	0.00	0.00	0.00
Random w ppl	0.870	21.28	6.89	1.53	12.65	12.65	8.00	2.00	6.00
Gittins	0.919	5.56	1.83	1.41	5.29	1.36	1.21	0.95	0.25
Greedy	1.167	2.33	1.46	1.26	2.22	1.09	0.00	0.00	0.00

Note. Policies are sorted by their performance (Optimality gap, column 2). Optimality gap denotes the distance between the best arm (in terms of value v_x) and chosen arm. Reversals are samples drawn from an arm that has been previously played and discarded in favor of another arm.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ambrus, A., and B. Greiner (2012), Imperfect Public Monitoring with Costly Punishment: An Experimental Study, *The American Economic Review*, *102*(7), 3317–3332.

Anderson, G. (1996), Nonparametric test of stochastic dominance in income distribution, *Econometrica*, *64*(5), 1183–1193.

Ariely, D., and K. Wertenbroch (2002), Procrastination, deadlines, and performance: self-control by precommitment., *Psychological science*, *13*(3), 219–24.

Athey, S., and G. Imbens (2016), Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360, doi:10.1073/pnas.1510489113.

Audibert, J., R. Munos, and C. Szepesvári (2008), Variance estimates and exploration function in multi-armed bandit, *Active Learning*.

Auer, P., N. Cesa-bianchi, and P. Fischer (2002), Finite time analysis of the multiarmed bandit problem, *Machine Learning*, *47*(2-3), 235–256, doi:10.1023/A:1013689704352.

Balafoutas, L., M. G. Kocher, L. Putterman, and M. Sutter (2013), Equality, equity and incentives: An experiment, *European Economic Review*, *60*, 32–51, doi:10.1016/j.euroecorev.2013.01.005.

Bao, J., and A. Wu (2017), Equality and Equity in Compensation, doi:10.1016/j.tree.2016.01.006.

Barber, A. E., and M. V. Roehling (1993), Job postings and the decision to interview: A verbal protocol analysis., *Journal of Applied Psychology*, *78*(5), 845–856, doi:10.1037/0021-9010.78.5.845.

Bardsley, N., and P. G. Moffatt (2007), The experimetrics of public goods: Inferring motivations from contributions, *Theory and Decision*, *62*(2), 161–193, doi:10.1007/s11238-006-9013-3.

Bell, G. D. (1969), Organizations: structure and behavior, John Wiley & Sons.

Belt, J. A., and J. G. Paolillo (1982), The Influence of Corporate Image and Specificity of Candidate Qualifications on Response to Recruitment Advertisement, *Journal of Management*, *8*(1), 105–112.

Bertsimas, D., and A. J. Mersereau (2007), A Learning Approach for Interactive Marketing to a Customer Segment, *Operations Research*, *55*(6), 1120–1135, doi: 10.1287/opre.1070.0427.

Bhattacharya, S., V. Krishnan, and V. Mahajan (1998), Managing New Product Definition in Highly Dynamic Environment, *Management Science*, (44), 50–64.

Biazzo, S. (2009), Flexibility, Structuration, and Simultaneity in New Product Development, *Journal of Product Innovation Management*, *26*(3), 336–353, doi: 10.1111/j.1540-5885.2009.00662.x.

Bochet, O., T. Page, and L. Putterman (2006), Communication and punishment in voluntary contribution experiments, *Journal of Economic Behavior and Organization*, *60*(1), 11–26, doi:10.1016/j.jebo.2003.06.006.

Bolton, G. E., and A. Ockenfels (2000), ERC: A Theory of Equity, Reciprocity, and Competition, *The American Economic Review*, *90*(1), 166–193.

Bornstein, G., and O. Weisel (2010), Punishment, Cooperation, and Cheater Detection in Noisy Social Exchange, *Games*, *1*(1), 18–33, doi:10.3390/g1010018.

Boudreau, J., W. Hopp, J. O. Mcclain, and L. J. Thomas (2003), On the Interface Between Operations and Human Resources Management, *MSOM*, *5*(3), 179–202.

Breugst, N., H. Patzelt, and P. Rathgeber (2015), How should we divide the pie? Equity distribution and its impact on entrepreneurial teams, *Journal of Business Venturing*, *30*(1), 66–94, doi:10.1016/j.jbusvent.2014.07.006.

Bubeck, S., N. Cesa-Bianchi, et al. (2012), Regret analysis of stochastic and nonstochastic multi-armed bandit problems, *Foundations and Trends® in Machine Learning*, *5*(1), 1–122.

Busenitz, L. W., and J. B. Barney (1997), Differences between entrepreneurs and managers in large organizations: Biases and heuristics in strategic decision-making, *Journal of Business Venturing*, *12*(1), 9–30, doi:10.1016/S0883-9026(96)00003-1.

Cappelen, A. W., A. D. Hole, and E. Sorensen (2007), The Pluralism of Fairness Ideals: An Experimental Approach, *The American Economic Review*, *97*(3), 818–827.

Cappelen, A. W., E. Sørensen, and B. Tungodden (2010), Responsibility for what? Fairness and individual responsibility, *European Economic Review*, *54*(3), 429–441, doi:10.1016/j.euroecorev.2009.08.005.

Caro, F., and J. Gallien (2007), Dynamic assortment with demand learning for seasonal consumer goods, *Management Science*, *53*(2), 276–292.

Chick, S. E. (2006), Chapter 9 Subjective Probability and Bayesian Methodology, *Handbooks in Operations Research and Management Science*, *13*(C), 225–257, doi: 10.1016/S0927-0507(06)13009-1.

Chick, S. E., and K. Inoue (2001), New two-stage and sequential procedures for selecting the best simulated system, *Operations Research*, *49*(5), 732–743.

Chick, S. E., J. Branke, and C. Schmidt (2010), Sequential sampling to myopically maximize the expected value of information, *INFORMS Journal on Computing*, *22*(1), 71–80.

Choo, A. S. (2014), Defining problems fast and slow: The U-shaped effect of problem definition time on project duration, *Production and Operations Management*, *23*(8), 1462–1479, doi:10.1111/poms.12219.

Cohen, M. A., J. Eliashberg, and T.-H. Ho (1996), New Product Development: The performance and time-to-market tradeoff, *Management Science*, *42*(2), 173–186.

Cooper, R. G., S. G. Edgett, and E. J. Kleinschmidt (1997), Portfolio management in new product development: Lessons from the leaders, *Research Technology Management*, *5*(40), 16–28.

Cooper, R. G., S. J. Edgett, and E. J. Kleinschmidt (2006), Portfolio management for new product development.

Corgnet, B., A. Sutan, and R. F. Veszteg (2011), My teammate, myself and I: Experimental evidence on equity and equality norms, *Journal of Socio-Economics*, *40*(4), 347–355, doi:10.1016/j.socec.2010.09.005.

Croson, R., and Y. Ren (2013), Overconfidence in Newsvendor Orders: An Experimental Study, *Management Science*, *59*(11), 2502–2517, doi:doi:10.1287/mnsc.2013.1715.

Cuzick, J. (1985), A Wilcoxon-type Test for Trend, *Statistics in Medicine*, *4*, 543–547.

Dal Bó, P., A. Foster, and L. Putterman (2010), Institutions and Behavior: Experimental Evidence on the Effects of Democracy, *The American Economic Review*, *100*(December), 2205–2229, doi:10.1257/aer.100.5.2205.

Deci, E. L., and R. M. Ryan (1985), *Intrinsic motivation and self-determination in human behavior*, Plenum, New York.

Dennis, A. R., J. S. Valacich, T. Connolly, and B. E. Wynne (1996), Process Structuring in Electronic Brainstorming, *Information Systems Research*, *7*(2), 268–277, doi:10.1287/isre.7.2.268.

Dennis, A. R., J. E. Aronson, W. G. Heninger, and E. D. Walker (1999), Structuring Time and Task in Electronic Brainstorming, *MIS Quarterly*, *23*(1), 95–108.

Deutsch, M. (1975), Equity, equality and need: What determines which value will be used as the basis of distributive justice?, *Journal of Social Issues*, *31*, 137–149.

Dow, S. P., K. Heddleston, and S. R. Klemmer (2009), The efficacy of prototyping under time constraints, *Proceeding of the seventh ACM conference on Creativity and cognition - C&C '09*, p. 165, doi:10.1145/1640233.1640260.

Duncker, K. (1945), On Problem-Solving, *Psychological Monographs*, *58*(5).

Ederer, F., and G. Manso (2013), Is Pay for Performance Detrimental to Innovation?, *Management Science*, *1909*, 1–18.

Encinosa, W. E., M. Gaynor, and J. B. Rebitzer (2007), The sociology of groups and the economics of incentives: Theory and evidence on compensation systems, *Journal of Economic Behavior and Organization*, *62*(2), 187–214, doi:10.1016/j.jebo.2006.01.001.

Engel, C. (2014), Social preferences can make imperfect sanctions work: Evidence from a public good experiment, *Journal of Economic Behavior and Organization*, *108*, 343–353, doi:10.1016/j.jebo.2014.02.015.

Erat, S. (2012), Making the Best Even Better: How Idea Pool Structure can make the Top Ideas Exceptional (working paper).

Erat, S., and V. Krishnan (2012), Managing Delegated Search Over Design Spaces, *Management Science*, *58*(3), 606–623, doi:10.1287/mnsc.1110.1418.

Farrell, J., and S. Scotchmer (1988), Partnerships, *The Quarterly Journal of Economics*, (May), 279–297.

Fehr, E., L. Goette, and C. Zehnder (2009), A Behavioral Account of the Labor Market: The Role of Fairness Concerns, *Annual Review of Economics*, (1), 355–384, doi:10.1146/annurev.economics.050708.143217.

Finke, R. A., T. B. Ward, and S. M. Smith (1992), *Creative Cognition: Theory, Research and Applications*, vol. 5, Bradford Books, doi:10.1006/ccog.1996.0024.

Fischbacher, U. (2007), Z-Tree: Zurich toolbox for ready-made economic experiments, *Experimental Economics*, *10*(2), 171–178, doi:10.1007/s10683-006-9159-4.

Fishbach, A., R. Dhar, and Y. Zhang (2006), Subgoals as substitutes or complements: the role of goal accessibility., *Journal of personality and social psychology*, *91*(2), 232–242, doi:10.1037/0022-3514.91.2.232.

Forbes, D. P. (2005), Are some entrepreneurs more overconfident than others?, *Journal of Business Venturing*, *20*(5), 623–640, doi:10.1016/j.jbusvent.2004.05.001.

Frazier, P. I. (2009), Knowledge-Gradient Methods for Statistical Learning, *Dissertation*, (June).

Frey, B. S., and S. Meier (2004), American Economic Association Social Comparisons and Pro-Social Behavior : Testing " Conditional Cooperation " in a Field Experiment Published by : American Economic Association Stable URL : http://www.jstor.org/stable/3592843 Social Comparisons and Pro-s, *94*(5), 1717–1722.

Gans, N., G. Knox, and R. Croson (2007), Simple Models of Discrete Choice and Their Performance in Bandit Experiments, *Manufacturing & Service Operations Management*, *9*(4), 383–408, doi:10.1287/msom.1060.0130.

Garivier, A., and O. Cappe (2011), The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond, *JMLR: Workshop and Conference Proceedings*, *19*, 1–18.

Gersick, C. J. G. (1988), Time and transition in work teams: toward a new model of group development, *Academy of Management Journal*, *31*(1), 9–41, doi:10.2307/256496.

Gersick, C. J. G. (1989), Marking Time: Predictable Transitions in Task Groups., *Academy of Management Journal*, *32*(2), 274–309, doi:10.2307/256363.

Gersick, C. J. G. (1991), Change Theories : Revolutionary Exploration of the Punctuated Paradigm, *The Academy of Management Review*, *16*(1), 10–36, doi:10.5465/AMR.1991.4278988.

Girotra, K., C. Terwiesch, and K. T. Ulrich (2010), Idea Generation and the Quality of the Best Idea, *Management Science*, *56*(4), 591–605, doi:10.1287/mnsc.1090.1144.

Gittins, J. (1979), Bandit Processes and Dynamic Allocation Indices, *Journal of the Royal Statistical Society. Series B (Methodological)*, *41*(2), 148–177.

Gittins, J., and D. Jones (1979), A dynamic allocation index for the discounted multiarmed bandit problem, *Biometrika*, *66*(3), 561–565, doi:10.2307/2335176.

Gittins, J., K. Glazebrook, and R. Weber (2011), *ulti-armed bandit allocation indices.*, John Wiley & Sons.

Graham, P. (2004), *Hackers and Painters*, O'Reilly Media, Inc.

Grechenig, C., A. Nicklisch, and C. Thoni (2010), Punishment despite Reasonable Doubt A Public Goods Experiment with Uncertainty over Contributions, *Journal of Empirical Legal Studies*, *7*(4), 1–32, doi:10.1111/j.1740-1461.2010.01197.x.

Gupta, S. S., and K. J. Miescke (1994), Bayesian look-ahead one-stage sampling allocations for selecting the largest normal-mean, *Statistical Papers*, *35*(2), 169–177.

Gupta, S. S., and K. J. Miescke (1996), Bayesian Look Ahead One-Stage Sampling Allocations For Selection Of the Best Population, *Tech. rep.*, Technical report Nr 94-30.

Gürerk, Ö. (2013), Social learning increases the acceptance and the efficiency of punishment institutions in social dilemmas, *Journal of Economic Psychology*, *34*(October), 229–239, doi:10.1016/j.joep.2012.10.004.

Gürerk, Ö., B. Irlenbusch, and B. Rockenbach (2006), The Competitive Advantage of Sanctioning Insitutions, *Science*, *312*(August 2016), 108–11, doi:10.1126/science.1123633.

Gürerk, Ö., B. Irlenbusch, and B. Rockenbach (2009), Voting with feet: community choice in social dilemmas, *Uni Erfurt Working Paper*, (4643), 1–46.

Hackman, J. R., and G. R. Oldham (1980), *Work redesign*.

Hellmann, T. F., and N. Wasserman (2016), The First Deal: The Division of Founder Equity in New Ventures, *Management Science*, pp. 1–23, doi:10.1017/CBO9781107415324.004.

Herz, H., D. Schunk, and C. Zehnder (2014), How Do Judgmental Overconfidence and Overoptimism Shape Innovative Activity?, *Games and Economic Behavior*, *83*, 1–23, doi:10.1016/j.geb.2013.11.001.

Holm, S. (1979), A Simple Sequentially Rejective Multiple Test Procedure, *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Holt, C. A., and S. K. Laury (2002), Risk Aversion and Incentive Effects, *American Economic Review*, *92*(5), 1644–1655.

Iansiti, M. (1995), Shooting the Rapids: Managing Product Development in Turbulent Environments, *California Management Review*, *38*(1), 37–58, doi:10.1016/0737-6782(96)82485-4.

Iyengar, S. S., and M. R. Lepper (2000), When choice is demotivating: can one desire too much of a good thing?, *Journal of personality and social psychology*, *79*(6), 995–1006, doi:10.1037/0022-3514.79.6.995.

Jared, N. W. D. (2016), Timing of reward allocations, Doctoral dissertation, Natinal University of Singapore.

Kalyanaram, G., and V. Krishnan (1997), Deliberate Product Definition: Customizing the Product Definition Process, *Journal of Marketing Research*, *34*(2), 276–285.

Kaufmann, E., N. Korda, and R. Munos (2012), *Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis.*, 661 pp., doi:10.1007/978-3-642-34182-3.

Kim, S.-H., and B. L. Nelson (2006), Chapter 17 Selecting the Best System, *13*(06), 501–534, doi:10.1016/S0927-0507(06)13017-0.

Konow, J., T. Saijo, and K. Akai (2009), Morals and mores? Experimental evidence on equity and equality, *Working Paper*, (June).

Kornish, L. J., and K. T. Ulrich (2011), Opportunity Spaces in Innovation: Empirical Analysis of Large Samples of Ideas, *Management Science*, *57*(1), 107–128, doi: 10.1287/mnsc.1100.1247.

Krishnan, V., and K. T. Ulrich (2001), Product Development Decisions: A Review of the Literature, *Management Science*, *47*(1), 1–21, doi:10.1287/mnsc.47.1.1.10668.

Krishnan, V., S. D. Eppinger, and D. E. Whitney (1997), Model-Based Framework Product to Overlap Development Activities, *Management Science*, *43*(4), 437–451.

Kroll, M., B. Walters, and S. A. Le (2007), the Impact of Board Composition and Top Management Team Ownership Structure on Post-Ipo Performance in Young Entrepreneurial Firms., *Academy of Management Journal*, *50*(5), 1198–1216, doi: 10.2307/20159920.

Kullback, S., and R. Leibler (1951), On Information and Sufficiency, *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lai, T. L. (1987), Adaptive Treatment Allocation and the Multi-Armed Bandit Problem, *The Annals of Statistics*, *15*(3), 1091–1114.

Lai, T. L., and H. Robbins (1985), Asymptotically efficient adaptive allocation rules, *Advances in Applied Mathematics*, *6*(1), 4–22, doi:10.1016/0196-8858(85)90002-8.

Levene, H. (1960), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, 278–292 pp.

List, J. A., A. M. Shaikh, and Y. Xu (2016), Multiple Hypothesis Testing in Experimental Economics, *NBER Working Paper Series*, p. 23, doi:10.3386/w21875.

Loch, C. H. (2017), Creativity and Risk Taking Aren't Rational: Behavioral Operations in MOT, *Production and Operations Management*, *26*(4), 591–604, doi: 10.1111/poms.12666.

Loch, C. H., and C. Terwiesch (1999), Accelerating the Process of Engineering Change Orders : Capacity and Congestion Effects, *Production Innovation Management*, (February 1999), doi:10.1111/1540-5885.1620145.

Locke, E. A., and G. P. Latham (2002), Building a practically useful theory of goal setting and task motivation. A 35-year odyssey., *The American psychologist*, *57*(9), 705–717, doi:10.1037/0003-066X.57.9.705.

Luhan, W. J., O. Poulsen, and M. W. M. Roose (2013), Unstructured bargaining over an endogenously produced surplus and fairness ideals: an experiment.

Maccormack, A., R. Verganti, and M. Iansiti (2001), Developing Products on Internet Time: The Anatomy of a Flexible Development Process, *Management Science*, *47*(1), 133–150.

Maillard, O.-A., R. Munos, and G. Stoltz (2011), A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences, *JMLR: Workshop and Conference Proceedings*, *19*(1), 497–514.

Mansfield, E. (1988), The Speed and Cost of Industrial Innovation in Japan and the United States: External vs. Internal Technology, *Management Science*, *34*(10), 1157–1168.

Maxwell, J. A. (2012), *Qualitative research design: An interactive approach*, Sage.

McFadden, D. (1973), Conditional logit analysis of qualitative choice behavior, doi: 10.1108/eb028592.

Metrick, A., and A. Yasuda (2010), *Venture capital and the finance of innovation*, John Wiley & Sons.

Miles, M. B., and A. M. Huberman (1994), *Qualitative data analysis: An expanded sourcebook*, Sage.

Moffatt, P. G. (2016), *Experimetrics: Econometrics for Experimental Economics*, Palgrave Macmillan, London.

Moore, D. A., and P. J. Healy (2008), The trouble with overconfidence., *Psychological Review*, *115*(2), 502–517, doi:10.1037/0033-295X.115.2.502.

Moreau, C. P., and D. W. Dahl (2005), Designing the Solution: The Impact of Constraints on Consumers' Creativity, *Journal of Consumer Research*, *32*(1), 13–22.

Morgan, W. R., and J. Sawyer (1979), Equality , Equity , and Procedural Justice in Social Exchange, *Social Psychology*, *42*(1), 71–75.

Moyer, M. (2012), *Funding Your Business Without Funds*, 1–216 pp., CreateSpace.

Nachum, G. (2015), Milestone-Based Vesting For Startup Founders, *Tech Crunch*.

Ng, P., W.-K. Wong, and Z. Xiao (2011), Stochastic Dominance via Quantile Regression (working paper).

Niño-Mora, J. (2011), Computing a classic index for finite-horizon bandits, *INFORMS Journal on Computing*, *23*(2), 254–267, doi:10.1287/ijoc.1100.0398.

Ostrom, E., J. Walker, and R. O. Y. Gardner (1992), Covenants With and Without a Sword : Self-Governance is Possible, *The American Political Science Review*, *86*(2), 404–417.

Özer, Ö., and O. Uncu (2013), Competing on time: An integrated framework to optimize dynamic time-to-market and production decisions, *Production and Operations Management*, *22*(3), 473–488, doi:10.1111/j.1937-5956.2012.01413.x.

Papanastasiou, Y., K. Bimpikis, and N. Savva (2017), Crowdsourcing Exploration, *Management Science*, (September), mnsc.2016.2697, doi:10.1287/mnsc.2016.2697.

Parvan, K., H. Rahmandad, and A. Haghani (2015), Inter-phase feedbacks in construction projects, *Journal of Operations Management*, *39*, 48–62, doi:10.1016/j.jom.2015.07.005.

Pasmore, W. A. (1988), *Designing effective organizations: The sociotechnical perspective*.

Pfeffer, J., and N. Langton (1993), The Effect of Wage Dispersion on Satisfaction , Productivity , and Working Collaboratively : Evidence from College and University Faculty, *Administrative Science Quarterly*, *38*(3), 382–407.

Powell, W. B., and I. O. Ryzhov (2012), *Optimal Learning*, John Wiley & Sons.

Putterman, L., J. R. Tyran, and K. Kamei (2011), Public goods and voting on formal sanction schemes, *Journal of Public Economics*, *95*(9-10), 1213–1222, doi:10.1016/j.jpubeco.2011.05.001.

Rapoport, A., and W. T. Au (2001), Bonus and Penalty in Common Pool Resource Dilemmas under Uncertainty., *Organizational behavior and human decision processes*, *85*(1), 135–165, doi:10.1006/obhd.2000.2935.

Rockenbach, B., and I. Wolff (2016), Learning and Peer Effects Designing Institutions for Social Dilemmas Designing Institutions for Social Dilemmas, *German Economic Review*, *17*(104), 316–336, doi:10.1111/geer.12103.

Roth, A. E. (1995), Bargaining Experiments, in *Handbook of Experimental Economics*, edited by John Kagel and Alvin E. Roth, pp. 253–348, Princeton University Press.

Sahbaz, F. M. (2013), Study of the Principles of Distributive Justice in Entrepreneurial Teams, Ph.D. thesis.

Saldaña, J. (2011), *Fundamentals of qualitative research*, Oxford university press.

Sawyer, K. (2012), *Explaining Creativity*, 2 ed., 8–9 pp., Oxford University Press.

Schwartz, E. M., E. T. Bradlow, and P. S. Fader (2017), Customer Acquisition via Display Advertising Using Multi- Armed Bandit Experiments Customer Acquisition via Display Advertising Using Multi-Armed Bandit Experiments, (August).

Scott, S. L. (2010), A modern bayesian look at the multi-armed bandit, *Applied Stochastic Models in Business and Industry*, *26*(6), 639–658.

Sethi, R., and Z. Iqbal (2008), Stage-Gate Controls, Learning Failure, and Adverse Effect on Novel New Products, *Journal of Marketing*, *72*(January), 118–134.

Shalley, C. E., and J. E. Perry-Smith (2001), Effects of social-psychological factors on creative performance: the role of informational and controlling expected evaluation and modeling experience., *Organizational behavior and human decision processes*, *84*(1), 1–22, doi:10.1006/obhd.2000.2918.

Sousa, S. (2010), Cooperation and Punishment under Uncertain Enforcement.

Strauss, A. L. (1991), *Qualitative analysis for social scientists*, Cambridge University Press.

Sutter, M., S. Haigner, and M. G. Kocher (2010), Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations, *Review of Economic Studies*, *77*(4), 1540–1566, doi:10.1111/j.1467-937X.2010.00608.x.

Sutton, R. S., and A. G. Barto (1998), *Reinforcement Learning: An Introduction*, MIT Press, doi:10.1109/TNN.1998.712192.

Terwiesch, C., and C. H. Loch (1999), Managing the Process of Engineering Change Orders: The Case of the Climate Control System in Automobile Development, *Production Innovation Management*, *6782*(98).

Thomke, S. H. (1998), Managing Experimentation in the Design of New Products, *Management Science*, *44*(6), 743–762, doi:10.1287/mnsc.44.6.743.

Tong, J., and D. Feiler (2016), A behavioral model of forecasting: Naive statistics on mental samples, *Management Science*, (December), doi:10.1287/mnsc.2016.2537.

Tyran, J. R., and L. P. Feld (2006), Achieving compliance when legal sanctions are non-deterrent, *Scandinavian Journal of Economics*, *108*(1), 135–156, doi:10.1111/j.1467-9442.2006.00444.x.

Ulrich, K. T., and S. D. Eppinger (2011), *Product design and development*, McGraw-Hill Education; 5 edition.

Vakili, S., and Q. Zhao (2016), Risk-averse multi-armed bandit problems under mean-variance measure, *IEEE Journal of Selected Topics in Signal Processing*, *10*(6), 1093–1111.

Verganti, R. (1999), Planned Flexibility: Linking Anticipation and Reaction in PD projects, *Journal of Production Innovation Management*, *16*, 363–376.

Wasserman, N. (2012), *The founder's dilemmas*, Princeton University Press.

Webster, D. M., and A. W. Kruglanski (1994), Individual Differences in Need for Cognitive Closure, *Journal of Personality and Social Psychology*, *67*(6), 1049–1062.

Wheelwright, S. C., and K. B. Clark (1992), *Revolutionizing product development: quantum leaps in speed, efficiency, and quality*, Simon and Schuster.

Whittle, P. (1980), Multi-Armed Bandits and the Gittins Index, *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*(2), 143–149.

Yin, R. K. (2013), *Case study research: Design and methods*, Sage publications.

Yuce, P., and S. Highhouse (1998), Effects of Attribute Set Size and Pay Ambiguity on Reactions to 'Help Wanted' Advertisements, *Journal of Organizational Behavior*, *19*(4), 337–352.