Ideation-Execution Transition in Product Development: Experimental Analysis

Evgeny Kagan
Ross School of Business, University of Michigan, ekagan@umich.edu

Stephen Leider
Ross School of Business, University of Michigan, leider@umich.edu

William S. Lovejoy
Ross School of Business, University of Michigan, wlovejoy@umich.edu

Bringing a new product to market involves both a creative ideation stage, and an execution stage. When time-to-market constraints are binding it is an important question how to divide limited time between the two stages and who should make this decision. We introduce a laboratory experiment that closely resembles this setting: it features a product development task with an open design space, a downstream cost increase and two development stages. We show that performance is significantly worse when designers choose for themselves when to transition from ideation to execution and that decision control explains a large share of performance variation even after controlling for individual differences. How the time is allocated between ideation and execution does not affect mean performance, but later transition increases risk. One driver of poor design outcomes in the designer-initiated transition regime are delays in physical construction and testing of designs. We show that such delays can be prevented by "nudging" designers towards early prototyping. However, the most important performance driver is the lack of task structure in endogenous regimes, which can be remedied by demanding a concrete, performance-oriented deliverable prior to a transition.

Key words: New product development, Behavioral operations, Innovation

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 (Cohen et al. 1996, Ho et al. 2002).

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 information 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 performance 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.

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.

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.

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.

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) who both study 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.

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.

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. Our experimental task is a version of a challenge used at international creativity competitions among high schools and colleges. 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 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] Appendix A presents detailed demographics data.

[2] For the exact transcript of the experimental instructions see Appendix D. The full set of instructions including the description of all measures collected during the experiment is included in the online appendix.

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 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. In section 6 we will consider three additional treatments in which transitions were also endogenously determined 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.

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]

4. Experimental Results

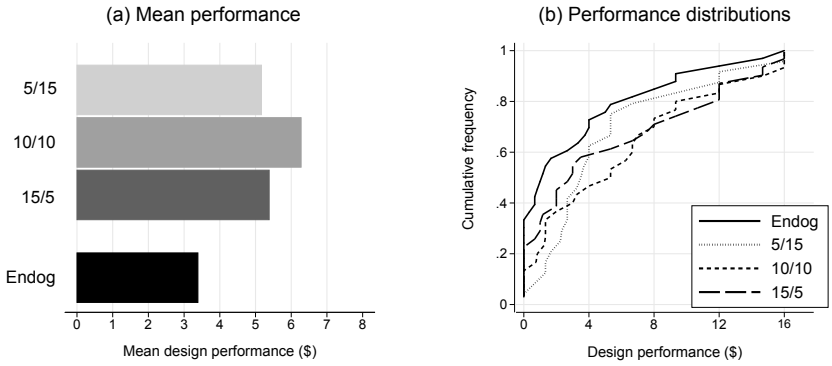
The remainder of this paper is organized as follows. This section will investigate whether design performance and design activities vary with the transition time and with the initiator of the

[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 was 20 minutes. Each subject was able to complete the task with the 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 results are not part of the analysis due to minor changes in the instructions, however including the pilot data does not affect the performance results reported in this section.

[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.

Figure 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.

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

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.

4.2. Performance comparisons: Results

4.2.1. Design performance comparisons Mean performance in each treatment is shown in panel (a) of Figure 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% if 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 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.

Regression results in Table 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 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 performance. Tobit regressions show similar effect magnitudes and slightly improved precision. 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

[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 C.

Table 1 Summary statistics of participant performance by treatment

	% Failures	Mean performance ($)	Mean performance given non-failure ($)
Endog	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

Note. % 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 shows mean performance of the subjects who were able to present a valid structure.

non-failure the treatment effects range from $1.77 to $2.50 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.

4.2.2. 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 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 (highest $p = 0.042$).[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 any given quantile, the size of the performance gap depends on the transition time and the quantile range (see figure 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)

[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 test based on splitting within-group outcomes into discrete categories and then comparing the incidence in each category between treatment groups. Different splits of the sample are possible as long as each category has a sufficient number of values. We find significant results ($p < 0.05$) for specifications with up to 4 categories. 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.

Table 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

Note. Probit, OLS and Tobit coefficients are reported. 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. * $p<0.1$, ** $p<0.05$, *** $p<0.01$.

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 ($p=0.069$ 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.

4.2.3. 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

[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 negligible as one goes from 10/10 to 15/5).

endogenous transitions, later transition times are associated with significantly reduced performance ($\rho = -0.365, p = 0.037$). 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 experiment 1, 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.

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

[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.

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 performance 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).

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.

4.4.1. 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 online appendix for our final coding scheme and figure 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.

Figure 2 A sample of design ideas

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

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]

4.4.2. 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:

[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 online appendix provides several inter-coder reliability measures for the design strategy variables. Most of the variables show high levels of consistency.

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 card. The fourth attribute, use of materials can include folding (F), tearing (T) and piercing (P), or any combination of those elements. Figure 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 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 2).

4.4.3. 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 B.[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.

4.5. Design Process: Results

4.5.1. 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 a risk-return trade-off in ideation time. The design process data are consistent with those performance results. In particular, in the Exogenous treatments time-to-last stacking increases monotonically in transition time (Trend test, $p < 0.01$), and later stacking is associated with higher failure rates ($\rho = 0.16$ across treatments, $p = 0.025$). Exogenous transitions also exhibit some differences in times-to-last idea, a measure of the extent to which subjects engage in exploration of the design space. In particular, with later transition subjects spend more time before they commit to an idea (6:22 min in 5/15, 07:34 min in 10/10, 08:05 min in 15/5). Delaying commitment to a design idea may improve the chances of finding an exceptional design, but it also increases the risk of failure, both of which is reflected in the performance data. However, the increase of the times-to-last idea in transition time is not significant (Trend test: $p = 0.156$), indicating that there is some degree of endogeneity in how much time is spent on ideation even with exogenous transition.[13]

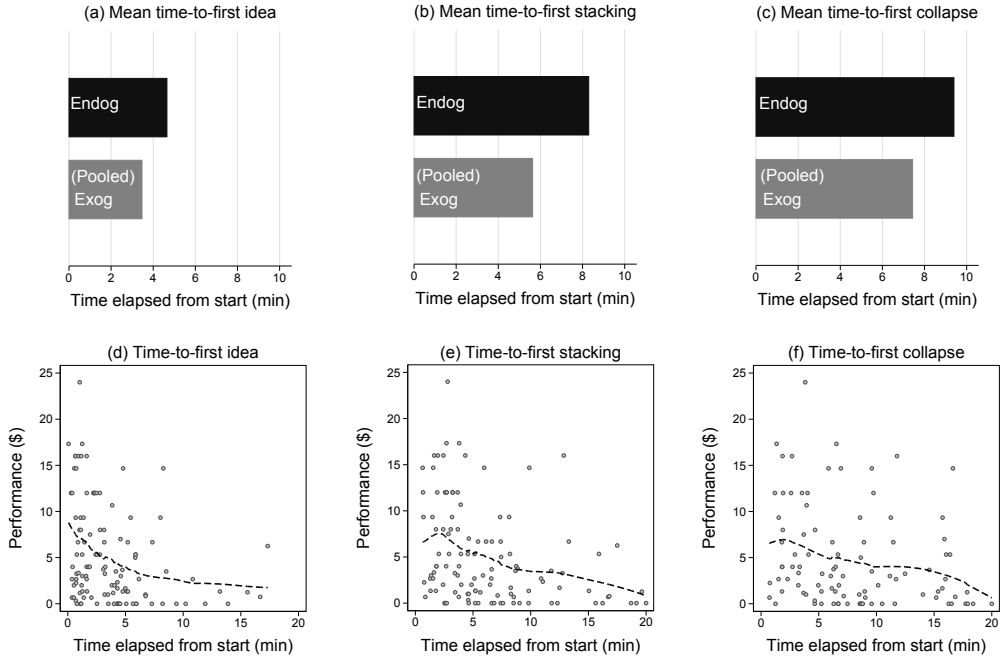
There are also several process differences between the (pooled) Exogenous and the Endogenous treatments. We focus in particular on the differences in activity timing.[14] Figure 3a)-3c) shows that the times of launching the first idea, the first test and experiencing the first collapse each occur with a substantial delay in the Endogenous treatment (Rank Sum tests, $p = 0.017$, $p = 0.011$, $p = 0.091$). That is, designers in the Endogenous condition spend more time pondering about possible design strategies or exploring the materials before launching a (recognizable) design idea. We also ran duration analysis (using Cox proportional hazard model) 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 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]

[13] Note that 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 ($07:34 < 10:00$, $p = 0.014$ and $08:05 < 15:00$, $p = 0.000$). 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 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 B for the complete list.

[15] 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

Figure 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).

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 3d)-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]

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 lower 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.

[16] To test for potential non-linearity 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 does 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.

Table 3 Relationships between performance and process variables

Dep.var.: Performance	(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	NA	17.78%	35.78%	11.23%

Note. Tobit coefficients are reported. 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". * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

4.5.2. 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 3 shows the Tobit coefficients and the percentages of common variation in performance explained by the process variables.[17]

[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 quadratic specifications, all of which confirmed our results.

Columns (2)-(4) of table 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 (Average marginal effect, $p < 0.01$) explaining about 36% of the performance variation. Time-to-first idea is also significantly related to performance explaining about 18% of performance variation ($p < 0.01$). The performance effect of first failure is somewhat weaker, explaining only about 11 % of the combined performance variation ($p = 0.033$). Note also 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.

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, and references there). 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.

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.

5.1. Experimental design

The basic setup of new treatments was similar to the four treatments of the main experiment: 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 online appendix). 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 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

[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.

a minimum viable construction (worth at least \$1, corresponding to the 25^{th} percentile of the performance distribution in experiment 1). Designers who were not able to demonstrate a minimum viable construction were not allowed to transition into execution receiving a payoff of \$0.[19]

5.2. Experimental results

5.2.1. 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, relative to the Endogenous base case (c.f. columns 1 and 2 of table 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 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 \$0.41 (\$0.34, \$0.30, 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.

[19] In all treatments (experiment 1 and 2) participants were paid based solely on their final performance; prototype performance was not incentivized.

[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.

5.2.2. 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 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, times-to-first idea and times-to-first stacking remain unchanged 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 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 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.

[21] When coding the video data from Experiment 2 we only recorded a subset of the original coding variables. The subset was selected based on the variables that were found to drive performance differences in experiment 1 (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 3 to 1.

[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 4 treatments (experiment 1) in table 3 and the full data set (experiment 1+2) in table 4.

Table 4 Experiment 2: Treatment comparisons and timing of activities

Dep Var: Performance	(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. 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 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. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

In sum, experiment 2 confirms 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.

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 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.

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.

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.[23]

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.

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.

[23] 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).

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

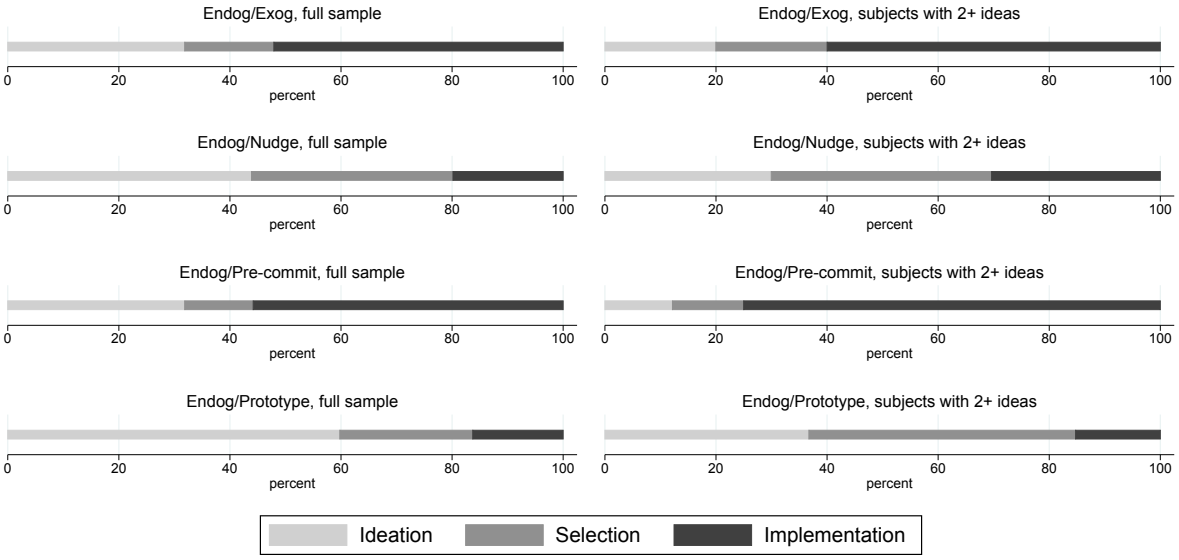
Next we investigate the extent to which the ability to generate ideas, to select an idea and to produce the best version of that idea drive performance differences between treatments. We use Rank Sum tests when making comparisons that involve the full sample of subjects, as well as OLS regressions, particularly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

6.2.1. 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 experiment 2. 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 ($p = 0.038$). 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]

[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

Figure 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.

The portions of the overall performance gap explained by the three metrics are summarized in the left half of figure 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

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: $p = 0.048$).

is made "on-the-go" (Nudge and Prototype) improve the quality of ideas and the ability to select good ideas, relative to the Endogenous base case.

6.2.2. 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).

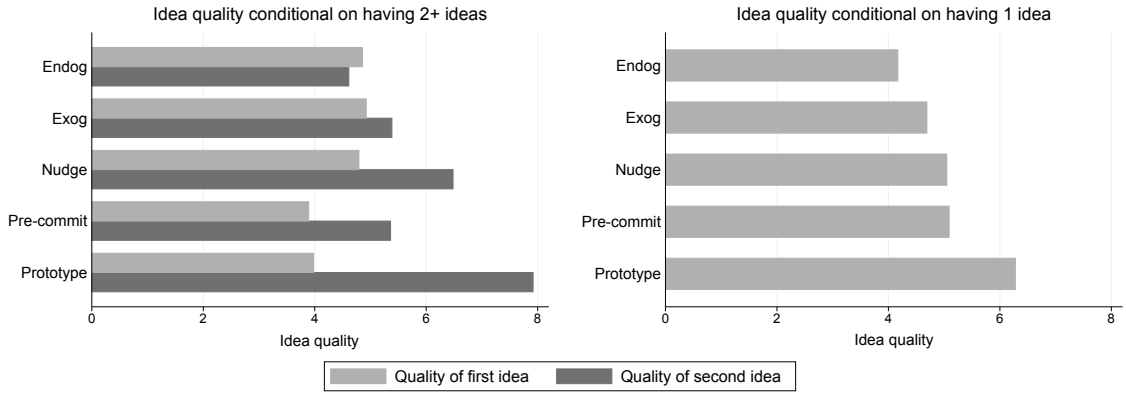
We find that restricting the sample to subjects with at least two ideas puts a greater weight on selection. Selection is driving a substantial portion of the performance gap between Endog and Nudge and of the gap between Endog and Prototype (37.93% and 43.84%; Rank Sum tests: $p = 0.155$ and $p < 0.01$, respectively). By contrast, selection explains no more than 20% of the performance advantage of Exog and Pre-commit. The right panel of figure 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 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 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 5 Idea quality by treatment



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.

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 microstructure 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 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]

[25] The data and the detailed description of the design projects are presented in the online appendix.

[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.

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.

Appendix

A. Participant demographics

Table 5 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

B. Main experiment, design activity variables

Table 6 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.

C. Multiple Hypothesis Adjustment

Table 7 Multiple hypothesis adjustment

Analysis	Variable	Coef	Unadjusted $p-$value	Adjusted $p-$value (Holm 1979)
Main experiment:	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
Main experiment:	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
Main experiment +	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
Main experiment +	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
Main experiment +	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
Main experiment +	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, and so on for the remaining adjustments.

D. 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 $\frac{9\times 8\times \$0.25}{3} = \$6$ 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 $\frac{9\times 8\times \$0.25}{3} = \$6$. 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.

References

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

Ariely, Dan, Klaus Wertenbroch. 2002. Procrastination, deadlines, and performance: self-control by precommitment. *Psychological science* **13**(3) 219–24. URL `http://www.ncbi.nlm.nih.gov/pubmed/12009041`.

Athey, Susan, Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113**(27) 7353–7360. doi:10.1073/pnas.1510489113. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1510489113`.

Bell, Gerald D. 1969. Organizations: structure and behavior. John Wiley & Sons.

Bhattacharya, Shantanu, Viswanathan Krishnan, Vijay Mahajan. 1998. Managing New Product Definition in Highly Dynamic Environment. *Management Science* (44) 50–64.

Biazzo, Stefano. 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. URL `http://doi.wiley.com/10.1111/j.1540-5885.2009.00662.x`.

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

Choo, Adrian 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, Morris A, Jehoshua Eliashberg, Teck-Hua Ho. 1996. New Product Development: The performance and time-to-market tradeoff. *Management Science* **42**(2) 173–186.

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

Cuzick, Jack. 1985. A Wilcoxon-type Test for Trend. *Statistics in Medicine* **4** 543–547. URL `http://onlinelibrary.wiley.com/doi/10.1002/path.1711680114/abstract`.

Deci, Edward L., Richard M. Ryan. 1985. *Intrinsic motivation and self-determination in human behavior*. Plenum, New York.

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

Dennis, Alan R., Joseph S. Valacich, Terry Connolly, Bayard E. Wynne. 1996. Process Structuring in Electronic Brainstorming. *Information Systems Research* **7**(2) 268–277. doi:10.1287/isre.7.2.268. URL `http://pubsonline.informs.org/doi/abs/10.1287/isre.7.2.268`.

Dow, Steven P., Kate Heddleston, Scott R. Klemmer. 2009. The efficacy of prototyping under time constraints. *Proceeding of the seventh ACM conference on Creativity and cognition - C&C '09* 165doi:10.1145/1640233.1640260. URL `http://portal.acm.org/citation.cfm?doid=1640233.1640260`.

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

Ederer, Florian, Gustavo Manso. 2013. Is Pay for Performance Detrimental to Innovation? *Management Science* **1909** 1–18.

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

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

Finke, Ronald A, Thomas B. Ward, Stephen M Smith. 1992. *Creative Cognition: Theory, Research and Applications*, vol. 5. Bradford Books. doi:10.1006/ccog.1996.0024.

Fishbach, Ayelet, Ravi Dhar, Ying 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.

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

Gersick, Connie 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, Karan, Christian Terwiesch, Karl T Ulrich. 2010. Idea Generation and the Quality of the Best Idea. *Management Science* **56**(4) 591–605. doi:10.1287/mnsc.1090.1144. URL `http://mansci.journal.informs.org/cgi/doi/10.1287/mnsc.1090.1144`.

Hackman, J. Richard, Greg R Oldham. 1980. *Work redesign*.

Herz, Holger, Daniel Schunk, Christian 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. URL `http://linkinghub.elsevier.com/retrieve/pii/S0899825613001474`.

Ho, Teck-Hua, Sergei Savin, Christian Terwiesch. 2002. Managing Demand and Sales Dynamics in New Product Diffusion Under Supply Constraint. *Management Science* **48**(March 2015) 187–206. doi:10.1287/mnsc.48.2.187.257.

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

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

Iansiti, Marco. 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. URL `http://search.ebscohost.com/login.aspx?direct=true{&}db=bth{&}AN=9512123298{&}site=ehost-live`.

Iyengar, Sheena S., Mark 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.

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

Kornish, Laura J., Karl 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, Viswanathan, Steven D Eppinger, Daniel E Whitney. 1997. Model-Based Framework Product to Overlap Development Activities. *Management Science* **43**(4) 437–451.

Krishnan, Viswanathan, Karl 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.

Levene, Howard. 1960. *Contributions to probability and statistics: Essays in honor of Harold Hotelling*.

List, John A., Azeem M. Shaikh, Yang Xu. 2016. Multiple Hypothesis Testing in Experimental Economics. *NBER Working Paper Series* 23doi:10.3386/w21875. URL `http://www.nber.org/papers/w21875`.

Loch, Christoph H, Christian 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, Edwin A, Gary 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.

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

Mansfield, Edwin. 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, Joseph A. 2012. *Qualitative research design: An interactive approach*. Sage.

Miles, Matthew B, A Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. Sage.

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

Ng, Pin, Wing-Keung Wong, Zhijie Xiao. 2011. Stochastic Dominance via Quantile Regression (working paper).

Özer, Özalp, Onur 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.

Parvan, Kiavash, Hazhir Rahmandad, Ali Haghani. 2015. Inter-phase feedbacks in construction projects. *Journal of Operations Management* **39** 48–62. doi:10.1016/j.jom.2015.07.005. URL `http://www.sciencedirect.com/science/article/pii/S0272696315000637`.

Pasmore, William A. 1988. *Designing effective organizations: The sociotechnical perspective*.

Saldaña, Johnny. 2011. *Fundamentals of qualitative research*. Oxford university press.

Sawyer, Keith. 2012. *Explaining Creativity*. 2nd ed. Oxford University Press.

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

Shalley, Christina E, Jill 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. URL `http://www.ncbi.nlm.nih.gov/pubmed/11162295`.

Strauss, Anselm L. 1991. *Qualitative analysis for social scientists*. Cambridge University Press.

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

Ulrich, Karl T, Steven D Eppinger. 2011. *Product design and development*. McGraw-Hill Education; 5 edition.

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

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

Wheelwright, Steven C, Kim B Clark. 1992. *Revolutionizing product development: quantum leaps in speed, efficiency, and quality*. Simon and Schuster.

Yin, Robert K. 2013. *Case study research: Design and methods*. Sage publications.