Making the Best Idea Better: The Role of Idea Pool Structure

January 22, 2017

Abstract

In any pool of ideas that is developed in a multistage process, it is often the case that only a small fraction of top ideas are exceptional. This study formulates a model of multistage development of idea pools, and examines the impact of the "structure" of idea pools and of learning across stages on the extent to which the top ideas in the pool are exceptional. Our key descriptive results demonstrate that the dispersion of ideas in the pool (defined in terms of the dis-similarity between ideas) has a positive impact on the value derived from the top ideas, especially so when learning across stages is limited. Subsequently, the study examines the normative question of optimal design of idea pools, and managerial rules-of-thumb are derived for the optimal choice of the number and dispersion of ideas within the pool.

1 Introduction

The new product development (NPD) funnel represents the classic view that NPD proceeds in stages where many ideas are pursued in earlier stages, allowing learning which enables the firm to winnow them down, to then develop into a few high-potential products that are ultimately launched. Within such a staged NPD process are the the stages of opportunity identification (and idea generation), concept development and selection, detailed design and engineering, testing, and launch (see for instance, Krishnan and Ulrich, 2001; Ulrich et al., 1995). While the initial stage is the fuzzy front-end that focuses on generating the raw set of ideas, the later stages focus on acquiring information to filter out the low quality ideas and develop only the better ideas.

Building off this classic conceptualization, Terwiesch and Ulrich (2009) employ the metaphor of a tournament to argue that an effective NPD process starts off with a pool of raw ideas, and then works to filter and narrow this pool by *pitting ideas against each other* and allowing only the ideas with better potential to proceed to the next stage. To illustrate with an example from Terwiesch and Ulrich (2009): Warner Brothers, the movie studio, identified/generated thousands of pitches for movies (most of them "duds"), developed hundreds of scripts, launched a few movies, out of which one was "Harry Potter," a movie that generated over $1 billion in revenues. Thus, in this view, the NPD process creates value for the firm by first identifying/generating a raw pool of ideas and then by selecting and developing the top ideas from this pool.

This article is primarily concerned with optimal design of the fuzzy front-end of an innovation process, i.e., the question of how many ideas to include and whether/when to allow multiple "similar" ideas in a raw idea pool. We illustrate our main question in the example below.

Suppose that a firm has access to three raw ideas, and wants to create an idea pool comprised of only two ideas to feed into their development process, with the goal of finally launching only the one idea turns out to be the best.[1] Since past research says that the quality of the raw idea (in the initial idea pool) is a key variable that determines the final value from the resulting innovation (Kornish and Ulrich, 2014), the firm should show preference for choosing those two ideas that have the highest ex-ante expected value (quality). Besides this intuitive driver, past literature also identifies

[1]For the sake of this example, the size of the idea pool (i.e., 2) is not viewed as a decision variable. However, note that past literature has examined how the size of the idea pool has a positive effect on the value of the top ideas, see for instance Dahan and Mendelson (2001).

a subtle role for the experimenter's uncertainty. Specifically, it argues that "[since] uncertainty arises from [the experimenter's] imperfect information about customers and markets, undiscovered or untested product designs and technologies, and challenges of perfectly executing and delivering even the ideal design" (Dahan and Mendelson, 2001, page 102), firms should "advance [or select] those ideas that have high variance in their [quality] ratings, that is, ones that some raters love but others hate" (Terwiesch and Ulrich, 2009, page 74). This uncertainty effect (Boudreau et al., 2011) requires that we select those two ideas that have the high degree of uncertainty.

Beyond these recommendations (mentioned above) that rely on *characteristics of a specific idea* (namely, the experimenter's ex-ante assessment of the mean and variance of the quality of a given idea), not much is known about how the choice should depend on the firm's ability to learn in multiple-stages, or about the determinants of the "optimal idea pool structure." Specifically, how does the value of the top idea change if we were to select two ideas for the pool that are similar in terms of some or many of their attributes? And how does this idea similarity interact with the experimenter's uncertainty regarding individual idea qualities, or with her ability to learn the "true quality" over multiple stages?

In the above example, we have abstracted away any consideration of where the three ideas (from which we select two) come from and why some ideas may end up being similar to others. Indeed, one may argue (as done by Afuah and Tucci, 2012) that a firm which employs broad search (for instance, crowdsourcing) is more likely to obtain ideas that are more dispersed (i.e., dis-similar to each other) compared to a firm that employs a local search-based (Kaizen-type) strategy. Thus, while we highlighted the role of selection as determining the idea pool structure, it should be noted that firms have other means (including the choice of whether or not to pursue crowdsourcing) to influence the structure of the idea pool. For the purposes of this article, we shall restrict ourselves to discussing and characterizing the optimal idea pool structure without referring to any explicit operational means that firms employ to achieve it.

Toward developing a formal model of general idea pools, we conceptualize the ideas as points in an abstract space of ideas (where the dimensions of the idea space correspond to the attributes of idea). Within this framework, we view the idea pool as a specific set of points. This conceptualization makes clear the (one-to-one) mapping between every idea pool and a unique configuration of points in the idea space, and thus allows us to associate the *distance between the points in the idea*

space to the "dispersion" of the corresponding idea pool. Next, we formulate an analytical model that explicitly captures (i) how a firm learns from the outcomes in one stage to reduce uncertainty about future stages, and (ii) how a key property of idea pool structure - namely the "dispersion" or dissimilarity between ideas - affects the realized outcomes. Subsequently, through formal analysis and a simulation study, we offer a descriptive characterization of how the idea pool structure the and learning between stages influences the extent to which the top ideas are exceptional. Our results, consistent with previous literature, confirm that a firm may gain exceptional value from their top ideas through three levers identified in past literature - "quantity/size of the idea pool," "[ex-ante] mean of the quality of ideas" and "variance [uncertainty] in the quality of ideas" (Terwiesch and Ulrich, 2009, page 28-29).

More importantly, our results clarifies the role of a related, but distinct lever - "dispersion of ideas in the pool" - that firms may employ to obtain exceptional value from their NPD process. Specifically, our key analytical results demonstrate that when dispersion in the idea pool is larger, the firm obtains greater value from their top ideas. Our model also offers a tractable and explicit characterization of how learning between stages can positively influence the value from top ideas.

With an eye on offering practicing managers more concrete guidance, we next consider idea pools where ideas can grouped into distinct idea categories, and we examine through computational experiments how the distribution of ideas among these categories affect the value from the best ideas. Our simulation results confirm the main qualitative insights from the analytical model while also showing that the value of top ideas is larger when the distribution of ideas is more skewed toward the category which is (i) ex-ante more promising, (ii) ex-ante more uncertain, (iii) permits greater between-stage learning, (iv) has greater dispersion between ideas. Moreover, our results also point to a subtle moderating influence of learning and dispersion by demonstrating that a skewed idea pool dominates a balanced idea pool when the NPD process permits significant learning or when the dispersion in each category is high, whereas the reverse is true when NPD process has only limited ability to learn from initial outcomes, or when the dispersion within each category is low.

Our normative examination of the "optimal idea pool structure" reveal a fundamental tradeoff: By designing an idea pool where the individual ideas are highly dispersed the firm can make the best ideas (from their idea pool) even better, and thus improve the effectiveness of the idea

pool. However, our results also show that such dispersion can come at a cost by forgoing possible economies of scale. Thus, designing highly dispersed idea pools, while improving the effectiveness of the idea pool, has the potential to reduce the efficiency of the NPD process. It is the tension between these two forces - efficiency loss and effectiveness gain due to greater dispersion in the idea pool - that determines the optimal structure of the idea pool.

Finally, our results reveal how the optimal design of idea pool is affected by the innovation process' ability to learn from early stages and its cost structure. NPD processes that are able to learn effectively (from earlier stages) should increase size of the idea pool and attempt to increase the dispersion between ideas. Moreover, high costs, either at the front-end or even in the later stages of the innovation process, while reducing the optimal number of ideas in the idea pool, also decreases the optimal dispersion that the firm creates in their idea pool. Thus, with a more costly development process, the firm benefits from optimally starting off with fewer, and more focused set of ideas.

2 Review of Related Literature

Our study is related to the literature that has examined the setting of product development and innovation through the lens of optimal search processes. In a seminal study, Dahan and Mendelson (2001) offer a formal parallel search model of concept testing, where the firm balances the costs incurred through testing against the additional benefits offered by searching through larger pools of concepts. This formulation, which has been followed by many subsequent product development studies, offers insights into how the cost of experimentation and the (probability) distribution of possible outcomes determine the optimal amount of experimentation.

Erat and Kavadias (2008) assume that testing one design allows one to also learn about related designs, and they use a model of correlation landscapes to extend Dahan and Mendelson's conceptual model to the case of sequential testing. They find that when learning between related designs is feasible, the overall test schedule splits into an initial exploratory phase (where testing occurs for the purpose of learning about designs) followed by an exploitation phase (where the firm quickly focuses in on the design that works).

Loch, Terwiesch, and Thomke (2001), recognizing that both parallel and sequential testing are feasible alternatives for a firm, offer an interesting comparison between them as well as an analysis of the contingencies which drive the optimality of one versus the other. Their findings suggest the importance of costs and of the possibility of learning in determining the optimal choice of search strategies.

While studies in this stream use "search" as a conceptual framework for studying concept testing and idea development, most of them view each concept (or design) as independent of other concepts. Indeed, this perspective ensures that the 'size of the pool' is the only relevant decision made by firms in these studies. Our focus is complementary in that we view both the size of the idea pool and additional structural characteristics of the idea pool (including idea dispersion) as possible choices that may be made by the firm in designing the optimal idea pool. In addition, the current research, by considering a multistage process where learning from earlier outcomes is feasible, also offers a novel and explicit characterization of the value of learning.

The question of dispersion of ideas within a pool, while unexamined within the context of a single firm (at least to the best of our knowledge), has been studied in the somewhat different setting of open innovation. In these contexts, the focal firm assigns a problem to multiple agents with the goal of ex-post selecting the best suggested solution (see Terwiesch and Xu, 2008). While much of this literature has focused on the problem of incentivizing the agent's effort, a small subset of this literature has concerned itself with understanding how too much similarity between the agent's solutions/ideas can be prevented; i.e., how to incentivize adequate dispersion. For instance, Kornish and Ulrich (2011) consider the possibility that when multiple agents suggest ideas, some of these ideas might end up being redundant. Examing a large dataset of ideas, they find that while full redundancy is not very common, there are significant similarities between ideas generated by independent agents.

In a setting of design contests, Erat and Krishnan (2012) offer a formal analysis of the possibility of generating similar ideas from innovation tournaments. They find that winner-take-all incentive mechanisms (in tournaments) make it more likely that the submitted solutions are similar. The current study complements these past studies and offers a formal model to quantify the role of idea pool structure, and specifically of the dispersion of ideas.

3 A Correlation Landscape Model of Idea Pools

Consider a firm which has an idea pool composed of N ideas. We view these ideas as being represented by points in a large, and very possibly infinite, idea space. This abstract idea space is composed of all possible ideas, with dimensions of the space being characteristics of individual ideas, such as the set of underlying technologies it employs, the technical and market assumptions it relies on, etc (for a related abstraction, see Kornish and Ulrich, 2011).

Ex-ante, the quality of each idea i $(i = 1, \cdots N)$ is uncertain, and is represented by the random variable $\mathbf{Q}_i$. This uncertainty is resolved as the firm expends resources and works on the idea. To parsimoniously capture such a stage-gate model of gradual resolution of uncertainty, we assume that there are 2 stages (labelled a and b), and that a part of the uncertainty is resolved at the end of first stage a, and the remaining uncertainty is resolved at the end of second stage b. We model this by representing the overall idea quality $\mathbf{Q}_i$ as follows: Let $\mathbf{Q}_i = \mathbf{Q}_i^a + \mathbf{Q}_i^b$ $(i = 1, \cdots N)$, where $\mathbf{Q}_i^a$ and $\mathbf{Q}_i^b$ are random variables, outcomes of which are realized at the end of stage a and stage b respectively.

We assume that the random variables $\mathbf{Q}_i^k$ $(k = a, b;\ i = 1, \cdots N)$ are jointly normally distributed with mean μ_i^k and standard deviation σ_i^k. Let the correlation between $\mathbf{Q}_p^k$ and $\mathbf{Q}_q^k$ be denoted by θ_{pq}^k $(k = a, b;\ p = 1, \cdots, N;\ q = 1, \cdots, N)$. Furthermore, let λ_{pq} $(p = 1, \cdots, N;\ q = 1, \cdots, N)$ be the correlation between $\mathbf{Q}_p^a$ and $\mathbf{Q}_q^b$.[2]

The focal firm starts off with N ideas, expends effort and observes $\mathbf{Q}_i^a$ at the end of stage a. Subsequently, the firm allows RN ideas to proceed to stage b. That is, the firm has a pass-rate of $R\%$ for the first stage and allows only $R\%$ of the ideas to proceed to second stage. Additional effort is expended on these RN ideas at stage b, and $\mathbf{Q}_i^b$ is observed.

Previous literature from both Strategy and New Product Development have employed (conceptually) similar models to represent the distance between ideas as measured on the idea space (Levinthal, 1997; Rivkin and Siggelkow, 2002; Gavetti et al., 2005; Erat and Kavadias, 2008).

[2]Correlations across stages makes learning feasible. Specifically, the outcome at the first stage provides us with information about potential outcome from second stage. For a related, but much less general specification, that looks at such sequential learning, see Erat and Kavadias (2008).

We should also point out that we cannot choose arbitrary within and between stage correlations $\theta_{pq}^a, \theta_{pq}^b, \lambda_{pq}$ since the correlation matrix needs to be positive definite for it to represent a valid multivariate normal distribution. But for now, we shall implicitly assume that the correlation matrix is valid (i.e., it is positive definite). Later, we shall offer, for a special case, a simple closed-form expression for the condition required for valid correlation matrix.

Specifically, in these models, it is assumed that when two ideas are "closer," (as measured on the idea space), their realized qualities would also be "closer." For example, in a typical model of tunable NK landscapes, such as the one presented in Levinthal (1997), it may be shown that the correlation between realized "performances" between two points decays exponentially with the distance between the points, and in the limit goes to zero (Kauffman, 1993, p. 63).[3] Thus, we shall assume that $\theta^k_{pq} \geq 0$ $(k = a, b \; ; \; p = 1, \cdots, N \; ; \; q = 1, \cdots, N)$, and we shall interpret θ^k_{pq}, the correlation between $\mathbf{Q}^k_p$ and $\mathbf{Q}^k_q$, in terms of the distance between ideas p and q; with greater (lesser) distance implying lower (greater) correlation.

The correlation across stages λ_{pq} has an intuitive interpretation in terms of the amount of learning yielded by the first stage. Specifically, the ex-ante standard deviation of an idea q at the second stage b is σ^b_q (i.e., the standard deviation of $\mathbf{Q}^b_q$); however, once an idea p has been examined at the first stage and we have observed $\mathbf{Q}^a_p$, then the conditional standard deviation of the same idea q is given by $\sigma^b_q\sqrt{1-\lambda^2_{pq}}$, or stated differently, the decreases in uncertainty (variance) is given by $\left(\sigma^b_q\lambda_{pq}\right)^2$. Hence, we interpret λ_{pq} in terms of the ability of the NPD process to learn from first stage, with higher (lower) values representing greater (lesser) amount of learning.

The above model allows us to represent idea pool structure in a straightforward manner. Specifically, in any pool of ideas, one assigns higher correlation θ_{pq} between idea qualities of any pair of ideas when the "similarity" between ideas is greater. Since the notion of "similarity" between ideas (and their "closeness") is key to representing the idea pool structure, we note that past literature have developed approaches to evaluating similarity between ideas, while leaving the "exact definition of similarity .. unspecified" (page 13, Griffin and Hauser, 1993; see also Kornish and Ulrich, 2011). Thus, even without explicitly specifying the dimensions of the landscapes or a formal metric for measuring the distance on it, it is possible to employ expert raters to evaluate the "similarity" between ideas, and to thus obtain (at least approximately) the correlation matrix corresponding to the idea qualities.

To complete the model setup, we make the following assumptions about the cost of and the value from the NPD process: Let C_a represent the cost incurred by the firm to work on all the N idea in first stage a, and let C_b represent the cost incurred to move all the selected RN ideas

[3] In our conceptualization, the firm can manage the idea pool structure, and consequently, the correlation parameter is an endogenous variable, in contrast to the NK landscape literature where the correlation parameter is typically viewed as exogeneous (for a notable exception, see Levinthal and Warglien, 1999)

through second stage b. Finally, let the total value (from the idea pool) obtained by the firm be the sum of qualities of all ideas that the firm carried through the second stage. That is, if S is the set of RN ideas that the firm permitted into stage b, then the total value obtained by the firm is given by $\sum_{i \in S} \mathbf{Q}_i$. Thus, implicit in our model is the assumption that a firm needs to work on an idea in both stages to derive any benefit from it.

In the next section we offer a descriptive characterization of outcomes and overall values that a firm may obtain from the modeled innovation process.

4 Descriptive Results: Idea Pool Structure and the Value from Top Ideas

Assume, without loss of any generality, that $\mu_i^b = 0$.[4] Now consider the firm's actions immediately after the stage a when part of the uncertainty about the idea quality has been resolved. Specifically, let the random variables $\mathbf{Q}_i^a$ $(i = 1, \cdots N)$ have realized values q_i^a. With this information, the firm has to choose RN ideas (out of N ideas) to admit to stage b. Obviously, the total value realized from the RN ideas that are admitted to the second stage is given by $\sum_{i=1}^{N} 1_{\{i \text{ is admitted}\}} \left(q_i^a + \mathbf{Q}_i^b\right)$; i.e., the sum of qualities of all the ideas that are admitted to stage b. Hence, viewed at the end of first stage a, the expected total value realized by the RN ideas that are admitted to second stage b is given $E\left[\sum_{i=1}^{N} 1_{\{i \text{ is admitted}\}} \left(q_i^a + \mathbf{Q}_i^b\right)\right] = \sum_{i=1}^{N} 1_{\{i \text{ is admitted}\}} \left(q_i^a + E\left[\mathbf{Q}_i^b | \mathbf{Q}_1^a = q_1^a, \mathbf{Q}_2^a = q_2^a, \cdots\right]\right)$.

The above expression, which we wish to characterize, depends on a large number of model parameters in a non-trivial manner. Specifically, it depends on μ_i^a, σ_i^a, σ_i^b $(i = 1, \cdots, N)$ and λ_{pq}, θ_{pq}^a $(p = 1, \cdots, N\ ;\ q = 1, \cdots, N)$, a total of $O(N^2)$ parameters, and is relatively intractable. Hence, to obtain analytical characterization, we make the following additional three simplifying assumptions in this section: (i) *identical means and variances*: assume that $\mu_i^a = \mu$ and $\sigma_i^a = \sigma_a$, $\sigma_i^b = \sigma_b$, (ii) *identical distance between ideas*: assume that the correlation $\theta_{pq}^a = \theta$ if $p \neq q$, and $\theta_{pq}^a = 1$ otherwise, (iii) *learning between stages restricted to within an idea*: assume that $\lambda_{pq} = \lambda$ if $p = q$, and $\lambda_{pq} = 0$ otherwise.

[4] Specifically, this assumption entails no loss in generality since we may always center the random variable $\mathbf{Q}_i^b$ by "shifting" it down by μ_i^b at the same time shifting $\mathbf{Q}_i^a$ up by the same amount (i.e., by redefining $\mu_i^a = \mu_i^a + \mu_i^b$).

The last assumption restricts the type of learning from each stage. Specifically, we assume that the first stage outcome of an idea will only gives us information and allow us to learn about the 2nd stage outcome of the **same** idea (and not any other idea). Also, note that the above assumptions imply a particular covariance structure. Hence, we shall also make the additional technical assumption that $\lambda \leq \min\left\{\frac{\sigma_a}{\sigma_b}(1-\theta), \frac{\sigma_b}{\sigma_a}\right\}$ to guarantee that the resulting covariance matrix is valid (positive definite). Details are available in the appendix B.

Given the above assumptions, the expression $\left(q_i^a + E\left[\mathbf{Q}_i^b | \mathbf{Q}_1^a = q_1^a, \mathbf{Q}_2^a = q_2^a, \cdots\right]\right)$ simplifies to

$$\left(\mu + \left(1 + \frac{\lambda\sigma_b}{\sigma_a(1-\theta)}\right)(q_i^a - \mu)\right) - \frac{\lambda r}{(1-r)(Nr+1-r)}\sum_{k=1}^{N}(q_k^a - \mu) \tag{1}$$

Detailed proofs are given in the technical appendix. With this expression we obtain the expected total value realized by the RN ideas that are admitted to stage 2 as

$$\sum_{i=1}^{N} 1_{\{i \text{ is admitted}\}}\left(q_i^a + E\left[\mathbf{Q}_i^b | \mathbf{Q}_1^a = q_1^a, \cdots\right]\right) = \sum_{i=1}^{N} 1_{\{i \text{ is admitted}\}}\left\{\begin{array}{c}\left(\mu + \left(1 + \frac{\lambda\sigma_b}{\sigma_a(1-\theta)}\right)(q_i^a - \mu)\right) \\ -\frac{\lambda r}{(1-r)(Nr+1-r)}\sum_{k=1}^{N}(q_k^a - \mu)\end{array}\right\}$$

Thus, a firm which maximizes the total (expected) value should rank order q_i^a, the realized values at stage a, and admit the top RN to stage 2. Consequently, viewed prior to stage 1, the expected value from the innovation process (not counting the costs) is given by the expected value of the sum of the highest RN numbers q_i^a $(i = 1, \cdots, N)$. More formally, let $\mathbf{Q}_{(i)}^a$ be the order statistics of the set of random variables $\mathbf{Q}_i^a$ $(i = 1, \cdots, N)$; i.e., $\mathbf{Q}_{(j)}^a$ is the j^{th} smallest of the random variables $\mathbf{Q}_i^a$ $(i = 1, \cdots, N)$. Then, the expected value from the idea pool is given by

$$\begin{aligned} V &= E\left[\sum_{i=1}^{RN}\left\{\left(\mu + \left(1 + \frac{\lambda\sigma_b}{\sigma_a(1-\theta)}\right)\left(\mathbf{Q}_{(N-i+1)}^a - \mu\right)\right) - \frac{\lambda r}{(1-r)(Nr+1-r)}\sum_{k=1}^{N}(\mathbf{Q}_i^a - \mu)\right\}\right] \\ &= \begin{array}{l} E\left[\sum_{i=1}^{RN}\left\{\left(\mu + \left(1 + \frac{\lambda\sigma_b}{\sigma_a(1-\theta)}\right)\left(\mathbf{Q}_{(N-i+1)}^a - \mu\right)\right)\right\}\right] \\ +E\left[\sum_{i=1}^{RN}\left\{-\frac{\lambda r}{(1-r)(Nr+1-r)}\sum_{k=1}^{N}(\mathbf{Q}_i^a - \mu)\right\}\right] \end{array} \end{aligned}$$

Note that the second term is 0 (since the expectation of $E[(\mathbf{Q}_i^a - \mu)] = 0$). Hence,

$$V = NR\mu + \left(1 + \frac{\lambda\sigma_b}{\sigma_a(1-\theta)}\right)E\left[\sum_{i=1}^{RN}\left(\mathbf{Q}_{(N-i+1)}^a - \mu\right)\right] \tag{2}$$

The next proposition fully characterizes V.

Proposition 1. *The expected value from the idea pool*

$$V = NR\mu + \sigma_a\sqrt{1-\theta}\left(1+\frac{\lambda\sigma_b}{\sigma_a\left(1-\theta\right)}\right)\Psi\left(R,N\right) \tag{3}$$

where $\Psi\left(R,N\right)$ *non-negative increasing in* R *(for* $R<\frac{1}{2}$*), and in* N*.*

While simple to directly read off from equation (3), it is useful to note the following comparative statics: The top ideas in the pool yield a greater value on expectation when

(i) the number of ideas is in the pool N is larger, or

(ii) the mean quality of an idea μ is larger.

(iii) the variance of the quality of an idea σ_a is larger

When the firm starts off with a larger idea pool (larger N), the top ideas are more likely to be exceptional. Similarly, if the idea pool is comprised of ideas that are ex-ante more promising (larger μ), then the firm is likely to end up with exceptional ideas of very high quality. Lastly, the result also shows that the expected value from the top ideas is increasing in the variance of idea qualities σ_a. Indeed, these three results are consistent with past research that recommends that firms should look to "increase the quantity," "increase the average quality" and "increase the variance of quality" of ideas (Terwiesch and Ulrich, 2009, page 28-29; see also Kornish and Ulrich, 2014).

The proposition 1 also allows us to characterize the effect of dispersion (i.e., θ), between-stage learning (λ), and of second stage uncertainty (σ_b). These are summarized next.

Proposition 2. *The expected value from the idea pool* V *is increasing in* λ*, increasing in* σ_b *when* $\lambda>0$ *(unaffected by* σ_b *when* $\lambda=0$*), and is decreasing in* θ*.*

When between-stage learning is greater (high λ), then more of the future uncertainty gets resolved earlier, thus allowing us to more accurately select the best ideas, consequently increasing the expected value. Moreover, our results also show it is only when learning from first stage outcomes is feasible (i.e., $\lambda>0$) that second-stage variance (σ_b) is beneficial. Thus, our model, by allowing for learning across stages (through the between-stage correlation parameter λ), helps us refine our

understanding of the "uncertainty effect" (Terwiesch and Ulrich, 2009, page 28-29) by demonstrating that the higher expected value from top ideas is not because of higher uncertainty/variance *per se*, but is due to the amount of uncertainty that is resolved at the time of the selection.

Finally, as the dispersion of ideas in the pool gets larger (lower θ), the exceptional ideas generate even higher value. It is important to note that the dispersion effect emerges from a related, but distinct mechanism compared to the "variance effect" found by past research (Boudreau, Lacetera, and Lakhani, 2011). Specifically, the variance effect refers to the fact that the best idea has greater value when each individual idea quality has greater variance. In our model, the variance of the quality of a single idea i is given by $\sigma_a^2 + \sigma_b^2 + \lambda\sigma_a\sigma_b$, which is independent of θ, and hence does not explain the dependence of value of exceptional ideas on the dispersion parameter θ.

By taking a broader view of what "uncertainty" means in the context of an idea pool, we may observe that the uncertainty as represented by the *covariances* is the measure that is affected by the dispersion parameter. Thus, our model helps clarify that the "uncertainty effect" found in previous literature results from a combination of two primitives: the variance/uncertainty of individual idea qualities ***and*** the covariance/dispersion between idea qualities. While the former is a property of an individual idea that exists independent of the idea pool, the latter is an intrinsic property of the specific collection of ideas itself.

5 Descriptive Results: Extension & Simulation Study

In the previous section, we derived qualitative insights about how the structure of the idea pool affects the value from top ideas. This section shall attempt to enrich these insight by examining additional, easily measurable, structural characteristics of idea pools.

5.1 Idea Pools with Idea Categories

Suppose the idea pool is such that we can split the overall idea space into K different categories, where each category is a smaller local region of the larger idea space. These categories may represent different underlying technologies, customer segments, business segments, etc. Suppose that, out of the N ideas, n_j $(j = 1, \cdots, K)$ ideas belong to category j; i.e., these n_j ideas in category j employ similar underlying technologies, target similar customer or business segments, etc.

Since each category is a local region (of the larger idea space), the distance between ideas in a given category is lesser compared to the distance between ideas across categories. Or equivalently, in terms of the model presented in §3, the correlation between idea qualities in a given category is greater compared to the correlation between idea qualities across categories. We assume that these correlations are such that the correlation between ideas p and q

$$\theta_{pq} = \begin{cases} 0 & \text{if ideas } p \text{ and } q \text{ belong to different categories} \\ \theta_j & \text{if ideas } p \text{ and } q \text{ belong to the category } j \end{cases}$$

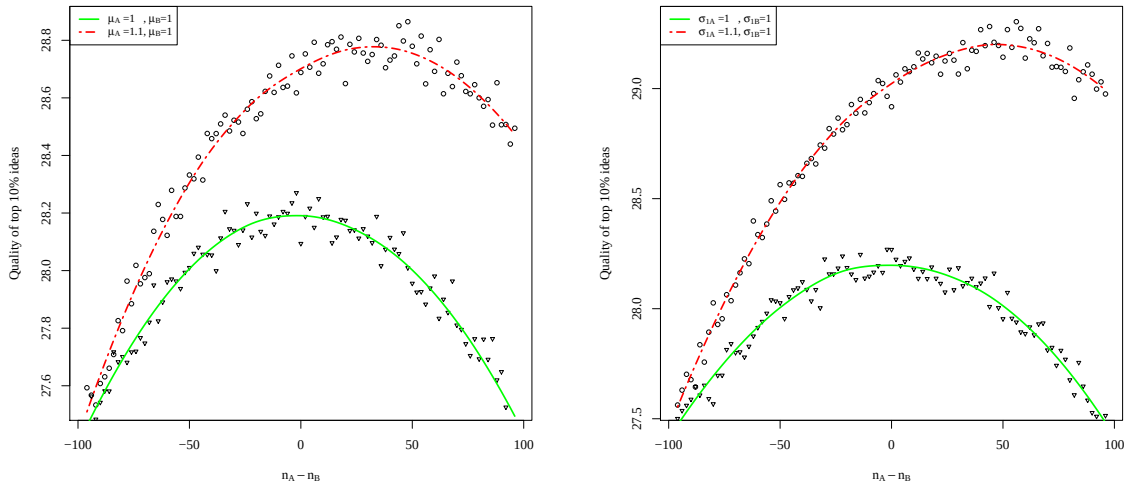
We assume that the learning is such that the between-stage correlation λ_{pq} between ideas p and q

$$\lambda_{pq} = \begin{cases} 0 & \text{if ideas } p \neq q \\ \lambda_j & \text{if } p = q \text{ and idea } p \text{ belong to the category } j \end{cases}$$

Lastly, as in the §4, we shall assume that the mean, and variance parameters within each category are identical. That is, μ_j is the mean quality, σ_{1j} is the first stage variance, and σ_{2j} is the second stage variance of every idea in category j. This model, while being a special case of the general model offered §3, still has a large number of parameters determining the "shape" of the idea pool (specifically $O(K)$ parameters). This makes systematic analysis and interpretation quite difficult. To overcome this problem, we shall first examine case of two categories A and B for a fixed idea pool size N (i.e., $n_A + n_B = N$), and subsequently demonstrate how the qualitative results extend to arbitrary K. To summarize the simulation setup, contingent on the total number of ideas N in the pool, the shape of the idea pool is fully determined by n_A (i.e., the number of ideas in category A), and the dispersion parameters θ_A and θ_B, the learning parameters λ_A and λ_B, means μ_A and μ_B, and standard deviations σ_{1A}, σ_{2A} and σ_{2B}, σ_{2B}.

For our base case, we fix the pool size $N = 100$, the mean values $\mu_A = \mu_B = 1$, its standard deviations $\sigma_{1A} = \sigma_{1B} = \sigma_{2A} = \sigma_{2B} = 1$, the correlation parameters $\theta_A = \theta_B = 0.2$, learning parameter $\lambda_A = \lambda_B = 0.1$. Subsequently, for each value of n_A (i.e., the shape of the idea pool), we use the statistical software R to simulate the outcomes from each pool 50,000 times to obtain V, the expected value of the sum of the top 10% of the ideas.[5] Subsequently, a curve is fitted (using

[5] Alternate values for several parameters, including for the pass rate R, were also checked and robustness of the



(a) Effect of mean

(b) Effect of uncertainty

Figure 1: Top idea quality as a function of $n_1 - n_2$

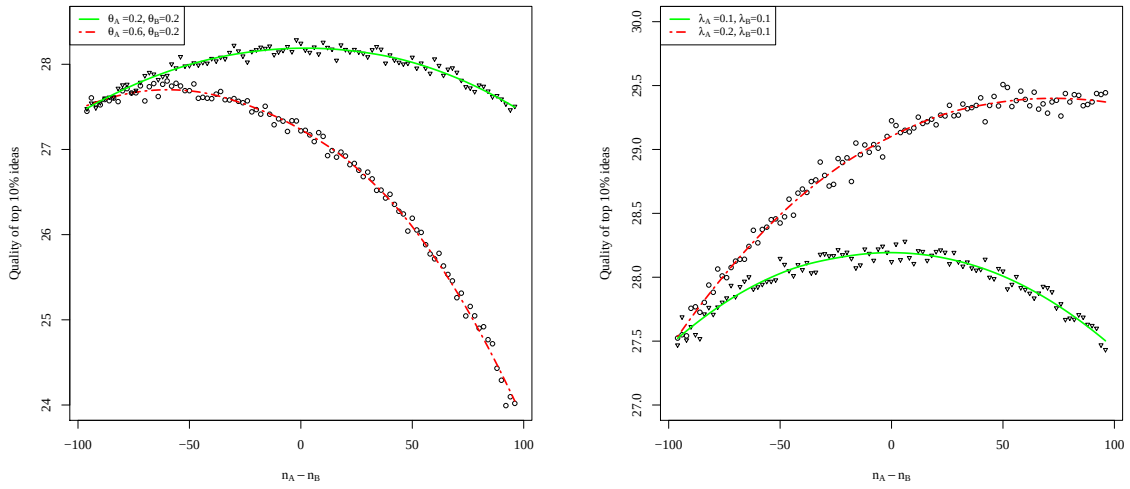
a LOESS regression, see Cleveland et al., 1988) for ease of visualization.

We shall interpret $n_A - n_B$ as the extent to which the idea pool is skewed toward category A, with a value of zero denoting a balanced idea pool. Figure 1a shows how the quality of the top ideas change as a function of $n_A - n_B$, for two settings, the base-case in which the categories are equally attractive $\mu_A = \mu_B = 1$ (solid green curve), and case in which category A is more attractive $\mu_A = 1.1$ and $\mu_B = 1$ (dashed red curve). As may be observed, if the two categories have identical ex-ante means, then the idea pool that is balanced and has equal number of ideas in each of the categories yields the highest value (i.e., the maximum of the solid green curve is when $n_A - n_B = 0$). However, the balanced idea pool no longer yields the highest V when the categories have different ex-ante means, and an idea pool which skews toward the more promising idea is superior (i.e., the maximum of the dashed red curve is when $n_A - n_B > 0$).

Similarly, Figure 1b shows how the shape of the idea pool affects value from top ideas when categories have different ex-ante uncertainty. As with the case of means, increase in the uncertainty in any category results in greater expected value. Moreover, idea pools which are skewed toward the category that is more uncertain yield greater value compared to idea pools that are balanced.[6]

qualitative results were confirmed. The code for running the simulation are available from the author.

[6] One of the reviewers pointed out that optimal idea pools will (almost) never be "corner solutions" with all ideas in the same category. Since V is concave in N, beyond a certain point, adding one more idea to less promising (or



(a) Effect of dispersion (θ)

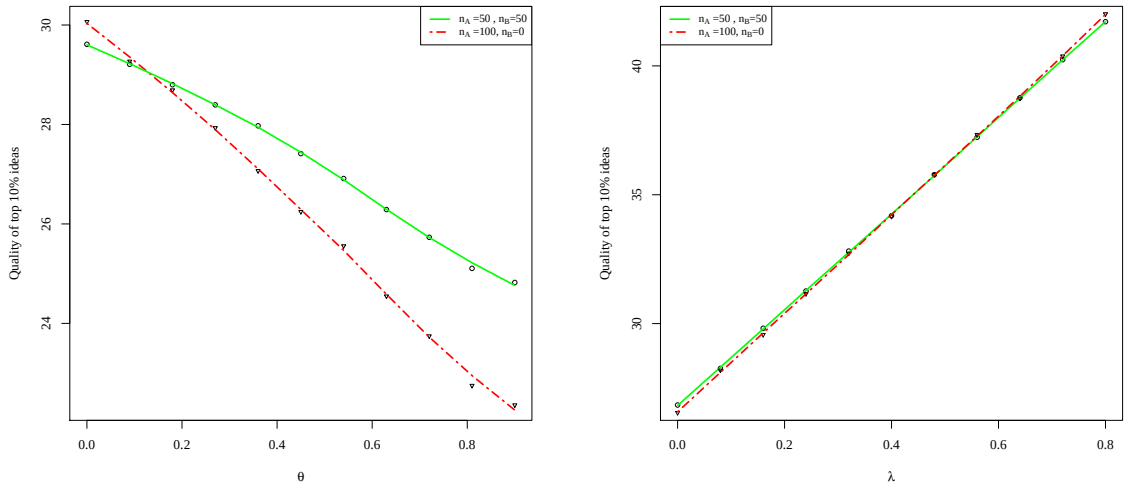
(b) Effect of learning (λ)

Figure 2: Quality of top ideas a function of θ for balanced idea pool (solid green) and skewed idea pool (dashed red)

Figure 2a shows how the value from top ideas change with dispersion parameter, and how this change depends on the shape of the idea pool. For a given idea pool, consistent with the insight from the simpler model, the expected value is decreasing when the dispersion within a (any) category decreases (i.e., the dashed red curve is always below the solid green curve). And, if the dispersion within any category decreases, then maximum value from the top ideas in a pool is not when we start out with a balanced idea pool, but when the idea pool skews toward the category that has greater within-category dispersion. It is interesting to note that this result has parallels in agent-based search models (such as Levinthal, 1997) which also find that broader search is beneficial when complexity of the landscape is higher. In our model, greater dispersion in any category (which is equivalent to greater complexity in these models) similarly results in the firm optimally focusing to devote more resources in this category.

Lastly, Figure 2b demonstrates that, for a given idea pool shape, the expected value is increasing when the learning is greater (i.e., dashed red curve is above the solid green curve). Moreover, the highest value is obtained by an idea pool which skews toward the category which allows greater learning.

less uncertain) category becomes superior to continuing to add to the more promising (or more uncertain) one.



(a) Effect of dispersion $(1-\theta)$

(b) Effect of learning (λ)

Figure 3: Quality of top ideas a function of θ for balanced idea pool (solid green) and skewed idea pool (dashed red)

The simulation above examined the effect of each parameter by itself. However, it is also of interest to know how the value from top ideas change (as a function of shape of idea pool) when the dispersion within every category changes(as can occur when the ruggedness of the corresponding idea landscape changes, for instance, see Levinthal, 1997; Kauffman, 1993), or when a superior NPD process allows greater the learning in every category. To examine this, we set $\theta_A = \theta_B = \theta$, and $\lambda_A = \lambda_B = \lambda$, and systematically varied θ and λ for the two extremes of idea pools (namely balanced pool $n_A = n_B$ and $n_A = 100$, $n_B = 0$).

Figures 3a and 3b show the value of top ideas as a function of the dispersion within every category and the learning-across-stages in every category. Consistent with our earlier results, the value of an idea pool increases with dispersion (lower θ) and increases with greater learning (greater λ). More interestingly, figure 3a shows that a balanced idea pool is more (less) beneficial compared to a skewed idea pool when the dispersion is smaller (larger). In contrast, a balanced idea pool is more (less) valuable compared to a skewed idea pool when the potential for learning is limited (large).

While we have only considered idea pools comprised of two categories so far, it is easy to see how these results generalize to more categories. Specifically, from an arbitrary number of categories,

we can always take any two categories and using the simulation results above demonstrate that (keeping all the other ideas in other categories the same) value from top ideas is higher if we skew the idea pool toward the category that has ex-ante higher mean, or has ex-ante greater uncertainty, or has ex-ante higher dispersion, or has ex-ante greater learning. That is, the highest value is obtained by an idea pool that has greater depth in any category that has higher ex-ante mean, ex-ante uncertainty, dispersion, and learning potential.

5.2 More general Idea Pool Structures

In §5.1 we considered one specific type of correlation structure corresponding to the notion that (some) idea spaces are possibly composed of distinct categories. Still, we might also be interested in more general contexts with arbitrary correlation structures. Analytical characterization of the value from top ideas, as may be imagined, is mathematically intractable. Hence, this section shall attempt to offer some preliminary answers by conducting an extensive simulation study that looks at how the broader correlation structure affects the value from top ideas.

Our analytical results (in the simpler case) demonstrated that the value from the top ideas depends on the correlation parameter, allowing us to conjecture that the value from top ideas (even in the case of arbitrary correlation structures) should possibly depend on some aggregate measure of dispersion. Hence, for the purpose of the simulation, for each correlation matrix, we define its dispersion[7] as

$$\text{DISP} = \sum_{i \neq j} (1 - \theta_{ij})$$

While DISP captures the total dispersion in the idea pool, it does not capture the extent to which the ideas in the pool are uniformly spread out rather than "clumping" together. As an example, consider the two correlation matrices shows in Table 1. The dispersion (DISP) for the left matrix is $6 \times (1 - 0.7) = 1.8$, and for the right matrix is $2 \times ((1 - 0.99) + (1 - 0.555) + (1 - 0.555)) = 1.8$. However, the matrix on the left represents a uniform and smooth distribution of ideas where the distance between any two ideas is identical, whereas the matrix on the left represents a distribution

[7] One may also use alternate measures of dispersion, such as DISP $= \Pi_{i \neq j} (1 - \theta_{ij})$ which instead of adding up individual dispersions multiplies them. Or alternatively, we could define it as DISP $= \sum_i (\Pi_j (1 - \theta_{ij}))$. We checked verified the robustness of our simulation results to such alternate definitions. In addition, our results also revealed that such alternate definitions add little in way of explanatory power above and beyond what is already captured by the simpler and more intuitive definition offered in the main text.

	idea 1	idea 2	idea 3
idea 1	1	0.7	0.7
idea 2	0.7	1	0.7
idea 3	0.7	0.7	1

	idea 1	idea 2	idea 3
idea 1	1	0.99	0.555
idea 2	0.99	1	0.555
idea 3	0.555	0.555	1

Table 1: Correlation matrices with same total dispersion, but different distributions

	Coefficient
Intercept	$2.615^{\ddagger}$ (0.022)
DISP	$0.003^{\ddagger}$ (0.00002)
UNIFORM	$-0.002^{\ddagger}$ (0.0002)

Table 2: Regression of V against DISP and UNIFORM

Note: $^{\ddagger}p < 0.001$

of ideas where ideas 1 and 2 are clumped together relative to idea 3. To parsimoniously capture this notion of uniformity of idea distribution, we shall define the uniformity of an idea pool as the negative of the square deviation of the individual dispersions, i.e.,

$$\text{UNIFORM} = -\sum_{i \neq j} \left((1 - \theta_{ij}) - \frac{\text{DISP}}{N\,(N-1)} \right)^2$$

Note that this measure is largest ($= 0$) when the idea pool is uniformly and smoothly distributed in the idea space, and is smallest (< 0) when individual ideas in the pool clump together as viewed on the idea space.[8]

For the simulation study, we set $N = 40$, $R = 10\%$, $\sigma_i^a = \sigma_i^b = 1$, $\mu_i = 0$, $\lambda = 0$ and generated 5000 random correlation matrices (using the approach offered in Joe, 2006, implemented by the clusterGeneration package in R). For each of these random matrices, we computed DISP and UNIFORM, and then simulated outcome qualities 10,000 times to calculate V (the value from the top 10% of ideas).

We conducted an OLS regression to systematically check how DISP and UNIFORM measures of a correlation matrix affect V. The results are shown in Table 2. As may be observed, our conjecture that increasing dispersion is related to increased value is supported in the regression.

Moreover, our results indicate that, keeping the dispersion fixed, lower values of UNIFORM

[8] While DISP measure representing averge dispersion is identical for the left and right matrices in Table 1, the UNIFORM measure representing clumping together of ideas is different. Specifically, for the left matrix, UNIFORM is $-6 \times \left((1-0.7) - \frac{1.8}{6}\right)^2 = 0$; and for the right matrix is $-2 \times \left(\left((1-0.99) - \frac{1.8}{6}\right)^2 + \left((1-0.55) - \frac{1.8}{6}\right)^2 + \left((1-0.55) - \frac{1.8}{6}\right)^2\right) = -0.26$.

measure (representing idea pools where ideas clump together) offer superior value. The intuition for this may be explained by going back to the two example correlation matrices we showed in Table 1. Note that the right matrix is at least as valuable as having only two ideas, but where the correlation is 0.55. Thus, when an idea pool is less uniform, it is equivalent to having fewer ideas which are less related to each other. This reduction in similarity between ideas, even at the cost of having fewer ideas in the pool, makes the top ideas more likely to be exceptional.[9] To summarize, our simulation study of more general idea pool configurations reveal that the top ideas are exceptional when the idea pool are highly, but not uniformly, dispersed.

So far in both §4 and §5, we have focused on offering descriptive results linking the idea pool structure to the value from the top ideas without any consideration of the cost side of the question. Next, we examine the normative implications of our results in §6.

6 Normative Implications: Design of the Idea Pool

For the analytical model given in §4, we make the following additional assumptions about the cost structure of the NPD process: Let the total cost of working on N ideas in the first stage, C_a, be composed of two parts. First, $c_1 g_1(N)$, where $c_1 \geq 0$, and $g_1(\cdot)$ a non-decreasing function, conceptually representing the (cost of) effort that would needed without any economies of scale. However, given that the ideas in the pool are not all dis-similar from each other, it appears plausible that there might be economies of scale and possible cost savings from working on similar ideas. For instance, if two ideas are very close to each other in the idea space (i.e., they are very similar), then the firm can possibly gain efficiencies in working on these two ideas. We shall assume that these cost savings can be represented by $h_1(\theta)$ where $h_1(\cdot)$ is an arbitrary non-decreasing function. Thus, the total cost of working on N ideas in the first stage $C_a = c_1 g_1(N) - h_1(\theta)$.

In the same manner, the total cost of working on RN ideas in the second stage is $C_b = c_2 g_2(RN) - h_2(\theta)$, where $c_2 \geq 0$, $g_2(\cdot)$ and $h_2(\cdot)$ are arbitrary non-decreasing functions. Finally, for ease of exposition, define $N_2 = RN$ as the number of ideas admitted to stage 2. With this, we

[9] A heuristic argument may be offered as follows: Recall from the main result (Proposition 1) that $V \propto \sqrt{1-\theta}\Psi(N) = \sqrt{1-0.77}\Psi(3) = 0.47\Psi(3)$ for the correlation matrix on the left side of Table 1 (UNIFORM=0). For the correlation matrix on the right side of Table 1 (UNIFORM < 0), V is at least as high as $\sqrt{1-\underline{\theta}}\Psi(N(\theta)) = \sqrt{1-0.55}\Psi(2) = 0.67\Psi(2)$. Note that while $\Psi(\cdot)$ is an increasing function, the term multiplying it is higher for the matrix which is less uniform, which drives the result that lower uniformity (keeping dispersion fixed) is related to greater value.

may write down the firm's net value from the idea pool as

$$\Pi\left(N, N_2; C_a, C_b, \theta\right) \quad = \quad V - \left(c_1 g_1\left(N\right) - h_1\left(\theta\right)\right) - \left(c_2 g_2\left(N_2\right) - h_2\left(\theta\right)\right)$$

where V is as given in Proposition 1. Thus, the firm's optimal choice of idea pool size N^*, and the optimal number of ideas admitted to second stage N_2^* are given by

$$\left(N^*, N_2^*\right) = \arg\max_{N, N_2} \Pi\left(N, N_2 \ ; \ C_a, C_b, \theta\right)$$

The next proposition characterizes the dependence of these optimal choices on the cost structures and on the dispersion of raw ideas in the idea pool.

Proposition 3. *N^* and N_2^* depends on $c_1, c_2, \mu, \sigma_1, \sigma_2, \lambda$ and θ as shown below.*

	c_1	c_2	μ	σ_a	σ_b	λ	θ
N^*	↘	↘	↗	↗	↗	↗	↘
N_2^*	↘	↘	↗	↗	↗	↗	↘

It is relatively intuitive why the optimal size of idea pool N^* should decline when the front-end costs c_1 are higher, just as it is relatively intuitive why the optimal number of ideas admitted to second stage N_2^* should decline when the later stage costs c_2 are higher. Specifically, in both cases, reducing the pool size or reducing the number allowed to second stage enables the firm to obtain an immediate savings in cost. But it is interesting that higher second stage costs c_2 also reduces the size of raw idea pool and higher front-end costs c_1 also reduces the number of ideas allowed to second stage. Intuitively, this occurs since higher later stage costs c_2, by reducing the number of ideas allowed into second stage, also make it less valuable to start off with larger raw idea pools. Similarly, higher front-end costs c_1, by reducing the value of large raw idea pools, also make it less valuable to have more ideas being allowed into second stage. The results make clear that determining the optimal idea pool size requires an understanding of the overall cost structure, and not just the initial front-end costs.

The proposition also demonstrates that the optimal size of the idea pool and the number of ideas admitted to second stage are both larger when the ideas are ex-ante more promising (higher μ), or when the idea qualities show higher ex-ante variability (higher σ_a or higher σ_b). The first

part is relatively intuitively - more promising ideas with higher μ will pay for themselves, and hence are worth adding to the idea pool, and taking forward to second stage. The intuition for the second part, namely dependence on σ_a and σ_b, is more subtle and mirrors the reasoning behind proposition 2. Specifically, we are able to learn from first stage outcomes (as long as $\lambda > 0$), and reduce uncertainty in second stage as soon as we examine the outcomes in the first stage. Consequently, the total resolved uncertainty in the first stage is related to both the first stage uncertainty σ_a and a fraction of second stage uncertainty σ_b. And similar to the logic of "uncertainty effect," when the total uncertainty is higher, the upside benefits are larger, especially so when the idea pool is larger. Hence, when more uncertainty is resolved at the end of first stage, the firm should optimally increase the size of the idea pool so as to gain these upside benefits.

The above argument also offers us the intuition as to why the size of the idea pool increases when λ increases. Specifically, with larger λ, the firm is able to learn more from the first stage results, consequently making the upside benefits from an idea pool greater, especially so when the idea pool is larger. Consequently, both N_1 and N_2 increase when the NPD process permits greater learning in the first stage. The proposition also shows that when the dispersion between ideas is lesser (i.e., θ is higher), then the firm should start off with a smaller idea pool, and should optimally allow fewer ideas into the second stage. Intuitively, with less dispersed ideas, ideas are more similar to each other, and consequently, reducing the pool size and forgoing some of these ideas allows the firm to gain from cost savings while at the same forgoing little by way of outcomes.

The above proposition treated the dispersion itself as an exogenous parameter. The next proposition endogenizes this and characterizes the firm's optimal choice of idea dispersion (and pool size).

Proposition 4. *N^*, R^*, and θ^* depends on $c_1, c_2, \mu, \sigma_a, \sigma_b, \lambda$ as shown below.*

	c_1	c_2	μ	σ_a	σ_b	λ
N^*	$\searrow$	$\searrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
N_2^*	$\searrow$	$\searrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
θ^*	$\nearrow$	$\nearrow$	$\searrow$	$\searrow$	$\searrow$	$\searrow$

The proposition shows that the optimal dispersion that the firm chooses in designing the idea pool is larger (smaller θ) when the ideas are more promising and/or when the idea quality is more variable. Intuitively, with more ex-ante promising ideas and/or ideas that are more variable in

quality, one wants to maintain the dispersion to maximize the chances of getting at least some exceptional idea.

With respect to the sensitivity to cost-structure, the proposition verifies that optimal size of idea pool and the optimal number of ideas allowed to second stage decrease when the costs increases, and that these choices increase when the ideas are ex-ante more promising or more variable. The proposition also shows that the firm should optimally have lesser dispersion when the costs are higher. Stated differently, when the costs are higher, the firm optimally starts off not only with with a smaller idea pool, but more interestingly, with a set of ideas that are more similar to each other! Intuitively, when costs are high, the idea pool size decreases (to allow for cost savings). Moreover, given the relatively high costs, the firm also saves through economies of scale by selecting more similar ideas.

Finally, with respect to the learning potential of the NPD process, the proposition shows that as the ability to learn from first stage outcomes become higher, the firm should optimally have larger and more dispersed idea pools. Intuitively, greater learning allow the firm to resolve more of the uncertainty, thus making it more valuable to have more and varied ideas.

These results offer a practicing manager useful guidance on the optimal design of an idea pool: When the development costs are higher and/or when the between-stage learning is limited and/or the ideas themselves are less promising/variable, the firm should optimally design the idea pool by limiting the number of allowable idea categories and by being more tolerant to clustering of ideas within a category.

7 Conclusions

How should a firm go about designing their idea pool so as to obtain exceptional value from their top ideas? Toward answering this question, the current article examines how the structural characteristics of the idea pool influence the value yielded by the top ideas in the pool. We conceptualize each idea as being represented by a point in a broader idea space, and develop an analytical model where the idea pool structure structure is linked to the joint distribution of ideas in the idea space. The model allows us to parsimoniously represent idea pools along a continuum - from 'dispersed' idea pools composed of dissimilar ideas, to more 'focused' idea pools comprising

of similar ideas clustered within fewer categories.

Analytical results and simulation studies provide convergent evidence that the structure of the idea pool can alter the extent to which the top few ideas are valuable. Specifically, our descriptive results show that the top ideas have exceptional value when the idea pool is skewed toward the category that has higher ex-ante mean, or higher ex-ante uncertainty, or higher within category dispersion, or greater between-stage learning potential.

Finally, given our interest in also offering managerial rules of thumb on optimal design of idea pools, we examined the normative implications of our descriptive results. Our analysis reveals that the optimal size of the raw idea pool one starts off with depends on the cost structure and the learning ability of the NPD process. When these costs are high (low) and/or when between-stage learning is low (high), the optimal idea pool design involves limiting (increasing) the number of allowable idea categories, and tolerating greater (limiting) clustering of ideas within a category.

References

Allan Afuah and Christopher L. Tucci. Crowdsourcing as a solution to distant search. *Academy of Management Review*, 37(3):355 – 375, 2012. ISSN 03637425.

Kevin J. Boudreau, Nicola Lacetera, and Karim R. Lakhani. Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science*, 57(5):843–863, 2011.

William S. Cleveland, Susan J. Devlin, and Eric Grosse. Regression by local fitting: Methods, properties, and computational algorithms. *Journal of Econometrics*, 37(1):87 – 114, 1988.

Ely Dahan and Haim Mendelson. An extreme-value model of concept testing. *Management Science*, 47(1):102–116, 2001.

Sanjiv Erat and Stylianos Kavadias. Sequential testing of product designs: Implications for learning. *Management Science*, 54(5):956–968, 2008.

Sanjiv Erat and Vish Krishnan. Managing delegated search over design spaces. *Management Science*, 58(3):606–623, 2012.

Giovanni Gavetti, Daniel A. Levinthal, and Jan W. Rivkin. Strategy making in novel and complex worlds: the power of analogy. *Strategic Management Journal*, 26(8):691–712, 2005.

Abbie Griffin and John R. Hauser. The voice of the customer. *Marketing Science*, 12(1):pp. 1–27, 1993.

Harry Joe. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177 – 2189, 2006. ISSN 0047-259X. doi: http://dx.doi.org/10.1016/j.jmva.2005.05.010. URL http://www.sciencedirect.com/science/article/pii/S0047259X05000886.

R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Number v. 1 in Applied Multivariate Statistical Analysis. Prentice Hall, 2002. ISBN 9780130925534. URL https://books.google.com/books?id=VlcZAQAAIAAJ.

Stuart A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, USA, 1 edition, 1993.

Laura J. Kornish and Karl T. Ulrich. Opportunity spaces in innovation: Empirical analysis of large samples of ideas. *Management Science*, 57(1):107–128, 2011.

Laura J. Kornish and Karl T Ulrich. The importance of the raw idea in innovation: Testing the sow's ear hypothesis. *Journal of Marketing Research*, 51:14–26, 2014.

Viswanathan Krishnan and Karl T Ulrich. Product development decisions: A review of the literature. *Management science*, 47(1):1–21, 2001.

Daniel A. Levinthal. Adaptation on rugged landscapes. *Management Science*, 43(7):pp. 934–950, 1997.

Daniel A. Levinthal and Massimo Warglien. Landscape design: Designing for local action in complex worlds. *Organization Science*, 10(3):342–357, 1999.

Christoph H. Loch, Christian Terwiesch, and Stefan Thomke. Parallel and sequential testing of design alternatives. *Management Science*, 47(5):663–678, 2001.

Weiliang Qiu and Harry Joe. *clusterGeneration: random cluster generation (with specified degree of separation)*, 2013. URL http://CRAN.R-project.org/package=clusterGeneration. R package version 1.3.1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/.

Jan W. Rivkin and Nicolaj Siggelkow. Organizational sticking points on nk landscapes. *Complexity*, 7(5):31–43, 2002.

Christian. Terwiesch and Karl Ulrich. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities*. Harvard Business School Press, Boston, 2009.

Christian Terwiesch and Yi Xu. Innovation contests, open innovation, and multiagent problem solving. *Management Science*, 54(9):1529–1543, 2008.

D.M. Topkis. *Supermodularity and complementarity*. Frontiers of economic research. Princeton University Press, 1998. ISBN 9780691032443.

Karl T Ulrich, Steven D Eppinger, et al. *Product design and development*, volume 384. McGraw-Hill New York, 1995.

A Main Appendix

Proof of Equation 1

The equation claimed that we may rewrite the expression $\left(q_i^a + E\left[\mathbf{Q}_i^b | \mathbf{Q}_1^a = q_1^a, \mathbf{Q}_2^a = q_2^a, \cdots\right]\right)$ in as

$$\left(\mu + \left(1 + \frac{\lambda\sigma_b}{\sigma_a\left(1-\theta\right)}\right)\left(q_i^a - \mu\right)\right) - \frac{\lambda r}{\left(1-r\right)\left(Nr+1-r\right)}\sum_{k=1}^{N}\left(q_k^a - \mu\right)$$

To prove claim, we first note that the distribution of$\mathbf{Q}_i^b$ conditional on the realized values $\mathbf{Q}_1^a = q_1^a, \mathbf{Q}_2^a = q_2^a, \cdots$ is normal (since $\mathbf{Q}_i^b; \mathbf{Q}_1^a, \mathbf{Q}_2^a, \cdots$) is multivariate normal. Hence, we may use standard results from multivariate statistics (for example, see Johnson and Wichern, 2002) to show that the conditional mean of $\mathbf{Q}_i^b$ is given by $\Omega_i \Sigma^{-1} \bar{Q}_1$ where

- Ω_i is an $1 \times N$ row vector with all elements equal to 0 except for its i^{th} element which is equal to $\lambda\sigma_a\sigma_b$;
- Σ is a symmetric $N \times N$ matrix with p, q^{th} element equal to $\left(\sigma_a\right)^2$ if $p = q$ and equal $\theta\left(\sigma_a\right)^2$ otherwise;
- and $\bar{Q}_1$ is the $N \times 1$ column vector with its p^{th} element given by given by $q_p^a - \mu$

Some algebra confirms the required claim**.**

□

Proof of Proposition 1

Recall that

$$V = NR\mu + \left(1 + \frac{\lambda\sigma_b}{\sigma_a\left(1-\theta\right)}\right) E\left[\sum_{i=1}^{RN}\left(\mathbf{Q}_{(N-i+1)}^a - \mu\right)\right]$$

Let $\hat{V} = E\left[\sum_{i=1}^{RN}\left(\mathbf{Q}_{(N-i+1)}^a - \mu\right)\right]$ where $\mathbf{Q}_{(j)}^a$ is the order statistics of $\mathbf{Q}_j^a$ $(j = 1, \cdots, N)$. Since the random variables $\mathbf{Q}_j^a$ $(j = 1, \cdots, N)$ are all of mean μ, standard deviation σ, and correlation θ, we may represent these random variables as follows: Define $\mathbf{Q}_j^a = \mu + \sigma_a\left(\sqrt{\theta}\mathbf{Z} + \sqrt{1-\theta}\mathbf{X}_j\right)$ where $\mathbf{Z}$ and $\mathbf{X}_j$ are standard normal random variables. Note that this representation is equivalent (in terms of distributions) since for the random variable $\mu + \sigma_a\left(\sqrt{\theta}\mathbf{Z} + \sqrt{1-\theta}\mathbf{X}_j\right)$, the mean is μ, the standard deviation is σ_a, and the correlation is $\frac{\sigma_a^2\theta}{\sigma_a\sigma_a} = \theta$.

Hence, we may rewrite V as follows:

$$\begin{aligned}
\hat{V} &= E\left[\sum_{i=1}^{RN}\left(\mathbf{Q}^{a}_{(N-i+1)}-\mu\right)\right] \\
&= E\left[\sum_{i=1}^{RN}\sigma_a\left(\sqrt{\theta}\mathbf{Z}+\sqrt{1-\theta}\mathbf{X}_{(N-i+1)}\right)\right] \\
&= E\left[\sum_{i=1}^{RN}\sigma_a\sqrt{\theta}\mathbf{Z}\right]+E\left[\sum_{i=1}^{RN}\left(\sigma_1\sqrt{1-\theta}\mathbf{X}_{(N-i+1)}\right)\right] \\
&= 0+\sigma_a\sqrt{1-\theta}E\left[\sum_{i=1}^{RN}\mathbf{X}_{(N-i+1)}\right] \text{ since } E\left[\mathbf{Z}\right]=0 \\
&= \sigma_a\sqrt{1-\theta}\Psi\left(R,N\right)
\end{aligned}$$

where $\Psi\left(R,N\right)=E\left[\sum_{i=1}^{RN}\mathbf{X}_{(N-i+1)}\right]$.

To show that $\Psi\left(R,N\right)$ is increasing in R and in N, it helps to define the function $\Gamma\left(N_2,N\right)=E\left[\sum_{i=1}^{N_2}\mathbf{X}_{(N-i+1)}\right]$. (i.e., $\Psi\left(R,N\right)=\Gamma(RN,N)$)

$\Gamma\left(N_2,N\right)$ is the expected value of the sum of highest N_2 numbers from a pool of N standard normal random variables. Hence, it is obviously the case that $\Gamma\left(N_2,N\right)$ is increasing in N (intuitively, for an arbitrary i, the i^{th} largest number chosen from $N+1$ numbers should be larger than the i^{th} largest number chosen from N numbers).[10]

To see that $\Gamma\left(N_2,N\right)$ is increasing in N_2, note that $\Gamma\left(N_2,N\right)-\Gamma\left(N_2-1,N\right)=E\left[\mathbf{X}_{(N-N_2+1)}\right]$. Moreover, since $\mathbf{X}_{(N-N_2+1)}$ is the order statistics, it obviously follows that $\mathbf{X}_{(N-N_2+1)}$ is decreasing in N_2.

Hence, $E\left[\mathbf{X}_{(N-N_2+1)}\right]\geq E\left[\mathbf{X}_{\left(N-\frac{N}{2}+1\right)}\right]$ for $N_2\leq\frac{N}{2}$. Moreover, since $\mathbf{X}$ is a standard normal random variable (with mean 0), $E\left[\mathbf{X}_{\left(N-\frac{N}{2}+1\right)}\right]$, being the expected value of the median is 0. Hence, $E\left[\mathbf{X}_{(N-N_2+1)}\right]\geq 0$ for $N_2\leq\frac{N}{2}$.

Which in turn implies that $\Gamma\left(N_2,N\right)$ is increasing in N_2 if $N_2\leq\frac{N}{2}$.

From the above two, it follows that $\Psi\left(R,N\right)$ $(=\Gamma(RN,N))$ is increasing in R and in N, as long as $RN\leq\frac{N}{2}$, i.e., $R\leq\frac{1}{2}$, and thus the proposition is proved.

Finally, since it is necessary for the proof of Proposition 3 and 4 later on, we note at this point itself that $\Gamma\left(N_2,N\right)-\Gamma\left(N_2-1,N\right)=E\left[\mathbf{X}_{(N-N_2+1)}\right]$ is increasing in N (since the N_2^{th} largest number chosen from N numbers is smaller than the N_2^{th} largest number chosen from $N+1$ numbers, as long as $k\leq\frac{N}{2}$). Thus, $\Gamma\left(N_2,N\right)$ is supermodular in (N_2,N). □

[10] A more formal proof that relies on showing that $\mathbf{X}_{((N+1)-i+1)}$ stochastically dominates $\mathbf{X}_{(N-i+1)}$ can be offered. We avoid doing so since this would require additional notation, and since the result is in any case quite intuitive.

Proof of Proposition 2

From proposition 1 we know that

$$V = NR\mu + \sigma_a\sqrt{1-\theta}\left(1+\frac{\lambda\sigma_b}{\sigma_a(1-\theta)}\right)\Psi(R,N)$$

Straightforward differentiation verifies that V is increasing in λ and σ_b.

To prove that V is decreasing in θ:

since $x+\frac{a}{x}$ is increasing when $x \geq \sqrt{a}$, it follows that $\sqrt{1-\theta}\left(1+\frac{\lambda\sigma_2}{\sigma_1(1-\theta)}\right) = \left(\sqrt{1-\theta}+\frac{\lambda\sigma_2}{\sigma_1\sqrt{1-\theta}}\right)$ is increasing in $1-\theta$ as long as $\sqrt{1-\theta} \geq \sqrt{\frac{\lambda\sigma_b}{\sigma_a}}$, i.e., as long as $\sigma_a(1-\theta) \geq \lambda\sigma_b$. That is, V is decreasing in θ as long as $\sigma_a(1-\theta) \geq \lambda\sigma_b$. Finally, to complete the proof we only need to observe that the constraint was assumed to hold to ensure that the parameters form a valid (positive definite) covariance matrix. □

Proof of Propositions 3 and 4

The firm maximizes the function

$$\Pi(N,N_2;C_a,C_b,\theta) = V-(c_1g_1(N)-h_1(\theta))-(c_2g_2(N_2)-h_2(\theta))$$

From the proof of Proposition 1,

$$V = RN\mu+\left(\sigma_a\sqrt{1-\theta}+\frac{\lambda\sigma_b}{\sqrt{1-\theta}}\right)\Gamma(N_2,N)$$

where $\Gamma(N_2,N) = E\left[\sum_{i=1}^{N_2}\mathbf{X}_{(N-i+1)}\right]$. Also, from the proof of Proposition 1, we know that $\Gamma(N_2,N)$ is increasing in N_2 and N, and that $\Gamma(N_2,N)$ is supermodular.

Hence, it follows that the function being maximized, $\Pi(N,N_2,c_1,c_2,\theta)$, is supermodular in $(N,N_2,-c_1,-c_2,-\theta,\mu,\sigma_a,\sigma_b,\lambda)$. This allows us to use standard monotone comparative statics results (for instance, see Topkis, 1998) to conclude that the maximizers N_2^*, N^* in Proposition 3 are increasing in $-c_1,-c_2,-\theta,\mu,\sigma_a,\sigma_b,\lambda$.

Similarly, the maximizers N_2^*, and N^* in Proposition 4 are increasing in $-c_1,-c_2,\mu,\sigma_a,\sigma_b,\lambda$, and the maximizer θ^* in Proposition 4 are decreasing in $-c_1,-c_2,\mu,\sigma_a,\sigma_b,\lambda$.

□

B Ancillary Appendix

Define the random variables $\mathbf{F_i} \sim \mathbf{N}(0, \lambda\sigma_a\sigma_b)$, $\mathbf{T_1} \sim \mathbf{N}\left(0, \theta\sigma_a^2\right)$, $\epsilon_{\mathbf{ia}} \sim \mathbf{N}\left(0, \sigma_a^2 - \theta\sigma_a^2 - \lambda\sigma_a\sigma_b\right)$, $\epsilon_{\mathbf{ib}} \sim \mathbf{N}\left(0, \sigma_b^2 - \lambda\sigma_a\sigma_a\right)$. Note that such random variables can be defined because the technical condition $\lambda \leq \min\left\{\frac{\sigma_a}{\sigma_b}(1-\theta), \frac{\sigma_b}{\sigma_a}\right\}$ implies that $\sigma_a^2 - \theta\sigma_a^2 - \lambda\sigma_a\sigma_b \geq 0$ and $\sigma_b^2 - \lambda\sigma_a\sigma_b \geq 0$. Moreover, we shall construct these random variables such that $\mathbf{F_i}$, $\mathbf{T_1}$ and $\epsilon_{\mathbf{ia}}$ are independent of each other and of $\epsilon_{\mathbf{ib}}$, and the $\epsilon_{\mathbf{ib}}$ is jointly multivariate normal for $i = 1, \cdots, N$ with correlation between $\epsilon_{\mathbf{ib}}$ and $\epsilon_{\mathbf{jb}}$ given by θ_{ij}^b.

Define $\mathbf{P_i^a} = \mathbf{F_i} + \mathbf{T_1} + \epsilon_{\mathbf{i1}}$ and $\mathbf{P_i^b} = \mathbf{F_i} + \epsilon_{\mathbf{ib}}$.

Then the joint distribution of $\left[\mathbf{P_1^a}, \mathbf{P_2^a}, \cdots, \mathbf{P_1^b}, \mathbf{P_2^b}, \cdots\right]$ is multivariate normal. Moreover, the variance of $\mathbf{P_i^a} = \sigma_a^2$, of $\mathbf{P_i^b} = \sigma_b^2$, the covariance $COV\left(\mathbf{P_i^a}, \mathbf{P_j^b}\right) = \lambda\sigma_a\sigma_b$ if $i = j$ and 0 otherwise, the covariance $COV\left(\mathbf{P_i^a}, \mathbf{P_j^a}\right) = \theta\sigma_a^2$ if $i \neq j$ (and σ_a^2 otherwise), the covariance $COV\left(\mathbf{P_i^b}, \mathbf{P_j^b}\right) = \theta_{ij}^b\sigma_b^2$ if $i \neq j$ (and σ_b^2 otherwise).

Thus, $\left[\mathbf{P_1^a}, \mathbf{P_2^a}, \cdots, \mathbf{P_1^b}, \mathbf{P_2^b}, \cdots\right]$ has identical distribution to the assumed joint distribution of $\left[\mathbf{Q_1^a}, \mathbf{Q_2^a}, \cdots, \mathbf{Q_1^b}, \mathbf{Q_2^b}, \cdots\right]$. Consequently, the assumed covariance matrix is valid (and thus positive definite) because of our explicit construction.