Idea Generation and the Quality of the Best Idea

Journal:	*Management Science*
Manuscript ID:	MS-01219-2007.R1
Manuscript Type:	R & D and Product Development
Keywords:	innovation, idea generation, extreme value theory, brainstorming, design

Idea Generation and the Quality of the Best Idea

Karan Girotra
Technology and Operations Management, INSEAD, Boulevard De Constance, 77305, Fontainebleau, France,
karan.girotra@insead.edu

Christian Terwiesch, Karl T. Ulrich
The Wharton School, University of Pennsylvania, 3730 Walnut Street, Philadelphia, Pennsylvania, 19104 US
terwiesch@wharton.upenn.edu, ulrich@wharton.upenn.edu

In a wide variety of organizational settings, teams generate a number of possible solutions to a problem, and then select a few for further investigation. We examine the effectiveness of two creative problem solving processes for such tasks— one, where the group works together as a team (the *team* process), and the other where individuals first work alone and then work together (the *hybrid* process). We define effectiveness as the quality of the *best* ideas identified by the group. We build theory that relates previously observed group behaviour to four different variables that characterize the creative problem solving process: (1) the average quality of ideas generated, (2) the number of ideas generated, (3) the variance in the quality of ideas generated, and (4) the ability of the group to discern the quality of the ideas. Prior research defines effectiveness as the quality of the average idea, ignoring any differences in variance and in the ability to discern the best ideas. In our experimental set-up, we find that groups employing the hybrid process are able to generate more ideas, to generate better ideas, and to better discern their best ideas compared to teams that rely purely on group work. Moreover, we find that the frequently recommended brainstorming technique of building on each other's ideas is counter-productive: teams exhibiting such build-up neither create more ideas nor are the ideas that build on previous ideas better.

Key words: Creativity, Brainstorming, Innovation, Idea Generation, Idea Selection, Team, Group, Individuals, Nominal Group, Interactive Build-up

History: This is the second version of this paper – June 11, 2009.

1 Introduction

Virtually all innovation processes include the creation or identification of opportunities and the selection of one or more of the most promising directions. When a movie studio creates a new feature film, it typically considers several hundred plot summaries, a few of which are selected for further development. When a company decides upon the branding and identity for a new product, it creates dozens or hundreds of alternatives, and picks the best of these for testing and refinement. When a consumer goods firm develops a new product, it typically considers many alternative concepts before selecting the few it will develop further. Generating the raw ideas that feed subsequent development processes thus plays a critical role in innovation.

The success of idea generation in innovation usually depends on the quality of the *best* opportunity identified. In most innovation settings, an organization would prefer 20 bad ideas and 1 outstanding idea to 21 merely good ideas. In the world of innovation, the extremes are what matter, not the average or the norm (Dahan and Mendelson (2001), Terwiesch and Loch (2004) Terwiesch and Ulrich (2009)). This objective is very different from those in, for example, manufacturing, where most firms would prefer to have 21 production runs with good quality over having 1 production run with exceptional quality followed by 20 production runs of scrap.

When generating ideas, a firm makes choices by intention or default about its creative problem solving process. In this paper, we investigate two commonly suggested organizational forms for idea generation. The first involves creation and evaluation of ideas by individuals working together as a *team* in the same time and space. The team approach is widely used in organizations (Sutton and Hargadon (1996)). Despite its wide usage, hundreds of experimental studies have criticized team processes as relatively ineffective (cf. Diehl and Stroebe (1987), Diehl and Stroebe (1991)). In the second approach, individuals work independently for some fraction of the allotted time, and then work together as a team. Such a *hybrid* process, also called the *nominal group* technique, has been suggested and studied in the prior literature as a way of effectively combining the merits of individual and team approaches. (cf. Robbins and Judge (2006), Paulus, Brown and Ortega (1996), Stroebe and Diehl (1994)). These studies find that the hybrid approach leads to more ideas and to higher satisfaction with the process among participants.

Notwithstanding its conflicting prescriptions, the existing brainstorming literature exhibits three gaps with respect to idea generation in innovation management. First, most papers focus on the number of ideas generated, as opposed to their quality, with the tacit assumption that *more* ideas will lead to *better* ideas. Second, the few papers that look at the quality of ideas look at the *average* quality of ideas as opposed to looking at the quality of the *best* ideas. Third, the focus of the existing literature is entirely on the *creation* process, and ignores the *selection* processes that teams apply to pick the most promising ideas for further exploration.

Given our focus on the use of idea generation in innovation, our metric for the effectiveness of the process is the quality of the ideas selected as the best. Building on prior work on innovation tournaments and on extreme value theory applied to innovation, we articulate a theory that combines the effects of four variables on the quality of the best idea: (1) the average quality of ideas, (2) the number of ideas generated, (3) the variance in the quality of ideas, and (4) the ability to discern the best ideas. Each of these variables affects the quality of the best ideas produced by a team or by a group employing the hybrid process.

We report on a laboratory experiment that compares the two idea generation processes with respect to each of these four variables individually and that measures their collective impact on the quality of the best idea. An accurate measurement of idea quality is central to our work. While most prior research has relied on the subjective evaluation of idea quality by one or two research assistants, we use two alternative approaches: a web-based quality evaluation tool that collects dozens of ratings per idea and a purchase intent survey that captures dozens of consumer opinions about their intention to purchase a product based on the idea. Our framework, with its emphasis on the importance of the best idea, and our novel experimental set-up let us make the following three contributions.

1. We find evidence that the best idea generated by a hybrid process is better than the best idea generated by a group process. This result is driven by the fact that the hybrid process generates about three times as many ideas per unit of time and that these ideas have significantly higher average quality.

2. We find that the hybrid process is better at identifying the best ideas from the set of ideas it previously generated. However, we also find that both group and hybrid processes are, in absolute terms, weak in their ability to discern the quality of ideas.

3. We show that idea generation in groups is more likely to lead to ideas that build on each other. However, in contrast to the common wisdom articulated by many proponents of group

brainstorming, we show that such build-up does not lead to better idea quality. In fact, we find that ideas that build on a previous idea are worse not better, on average.

The remainder of this paper is organized as follows. We review the relevant literature in Section 2. We then develop in Section 3 hypotheses about the differences between the team and hybrid processes with respect to these four main process variables. Section 4 describes the experiment. Section 5 reports how the organization of the idea generation process influences the variables determining the quality of the best idea. In Section 6, we examine how the effects of these variables come together by comparing the quality of the best idea across treatments. Section 7 looks at the micro-level data capturing the effects of idea build-up, and Section 8 contains concluding remarks.

2 Literature

The role of organizational processes in idea generation has been examined in the social psychology literature and in the innovation management literature. The social psychology literature has examined the idea generation process in detail, and is often called the *brainstorming* literature. The innovation management literature has focused on innovation outcomes and organizational forms.

The social psychology literature mostly originates with Osborne's 1957 book, *Applied Imagination* (Osborne (1957)), which introduces the term *brainstorming*. Osborne argued that working in teams leads to multiple creative stimuli and to interaction among participants, resulting in a highly effective process. His argument spawned many studies that tried to verify experimentally this argument. Diehl and Stroebe (1987) and Mullen, Johnson and Salas (1991) provide a detailed overview of this literature. These studies experimentally examined groups generating ideas as teams or as individuals. In terms of performance metrics, the literature focuses on the average quality of the ideas generated, the number of ideas generated, and measures that combined the two such as the total quality produced. Quality ratings for ideas generated are typically provided through evaluations by research assistants. For example, in Diehl, M., and W. Stroebe (1987), the ideas were rated by one research assistant and a second assistant was used to verify the reliability. The research has unequivocally found that the number of ideas generated (i.e., productivity) is significantly higher when individuals work by themselves and the average quality of ideas is no different between individual and team processes. (All of these studies normalize for total person-time invested to control for differences in the numbers of participants and the duration of the activity.) Thus, team processes have been found to be significantly inferior to individual processes. This main conclusion is in stark contrast with Osborne's hypothesis and to anecdotal evidence that team idea generation processes (i.e., brainstorming) are widely used in organizations.

In line with the social psychology literature we also conduct experiments. However, in contrast to this literature, we examine idea generation in the specific context of generating ideas in response to an innovation challenge. Given the focus on innovation, we are concerned with the quality of the *best* ideas resulting from the idea generation process, not with the average quality. Furthermore, we depart from this literature by employing a novel method of evaluating idea quality based on a large panel of independent raters and on a purchase-intent survey conducted with subjects from the target market segments.

To resolve the paradox between the social psychology literature and the anecdotal evidence, Sutton and Hargadon (1996) conducted a field-based observational study of the product design consulting firm IDEO. They found that contextual differences between the lab and the real world such as the nature of problems may explain the contrast between practice and the laboratory findings. More recently, Kavadias and Sommer (2007) take an innovative approach to this paradox. They show analytically that the specific nature of the problem and group diversity matters to the difference in the performance of individuals and teams. In particular, they conjecture that the experimental evidence may be an artifact of exploring simple idea generation problems which are not representative of real situations.

The role of organizational structure in the idea generation process has also been examined empirically, most notably, by Fleming and Singh (2007), who use patent data to study differences in productivity, quality, and quality variance between inventors who work by themselves and those who collaborate. Quality is measured as the number of citations received by the patent. Taylor and Greve (2006) examine average quality and variance of creative output in the comic book industry. The quality is measured using the collector-market value of a comic. While Fleming and Singh (2007) find that quality variance is higher for inventors working individually, Taylor and Greve (2006) find the opposite. In the experimental studies mentioned before, the differential resource investment between individuals and teams can be controlled by aggregating individual innovators into synthetic teams (also called *nominal groups*); this is impossible to do in natural empirical studies. Thus, it is hard to draw conclusions about productivity from these studies, though the results on average quality and variance directly inspire our work.

Lastly, the statistical view of innovation, which is at the core of our analyses and hypotheses was first developed by Dahan and Mendelson (2001). They model creation as a series of random draws from a distribution followed by a selection from the generated ideas. We employ this model to identify the statistical properties that influence the quality of the best idea. We summarize the relevant literature and the key differences between the literature and our study in Table 1.

3 A Theory of Creative Problem Solving

For simplicity, in this section we define quality as a single dimension of merit, although in testing our theory, we will consider multiple, alternative dimensions. Creative problem solving can be thought of as two steps: generating a pool of ideas (*idea generation)* and evaluating and selecting from this pool of ideas (*idea selection.*) For most problems, the quality of the ideas identified in the idea generation step is not objectively discernable. Thus, the problem solving entity usually makes a subjective estimate of the quality of each idea, and then selects a subset of the most promising ideas for further development. The subset is generally composed of ideas that have the highest subjective assessment of quality. Typically, the selected subset is substantially smaller than the original pool of ideas, and so the overall process exhibits a tournament structure (Terwiesch and Ulrich (2009)).

For the organization, the payoff from this process depends on the quality of this selected subset of ideas, and on the outcome of subsequent development activities and external events. Given our focus on the process of generating and selection ideas, we use the quality of the selected subset of ideas as the key performance measure. In this section, we build a theory that explains the causal relationships between the organizational processes employed in creative problem solving and the quality of the selected subset. We divide this theory into the two steps of generation and selection. The elements of the theory are summarized in Figure 1 and described below.

3.1 The Idea Generation Step

The quality of the selected subset of ideas depends on the pool of ideas available from which selection can be made. For most reasonable selection schemes, the quality of the selected ideas will be better if a superior pool of initial ideas is available. There are three process variables that can lead to a superior pool of ideas.

1. If the *mean quality* of the ideas created or identified initially is higher, the quality of the selected subset will also be higher.[1]

2. The *number of distinct ideas generated* also influences the quality of the selected subset. If an equal number of ideas, the *best n*, are selected from the initial pool, the *best n* from a larger pool will be better on average than the *best n* from a smaller pool. For example, the tallest 5 people from a city of 1,000,000 inhabitants will be taller than the tallest 5 people from a city of 1,000 inhabitants, assuming the same distributional characteristics of height in the two cities.

3. The *variance in quality* of the ideas in the pool also influences the quality of the selected subset. As an extreme example, consider two pools of ideas– one in which all ideas are of the same quality, say 5 on a 10 point scale; and the another pool with the same number of ideas but in which half of have quality 9 and the other half 1. These two pools are the same size and have the same mean quality. However, if we were to select the best idea from each of the pools, on average the idea selected from the second pool will be better. This logic extends to selection of the best-n ideas.[2]

Now we discuss how the choice of organizational process (team vs. hybrid) influences each of these process variables.

[1] Formal proofs for this and other statistical statements are provided in the Appendix.

[2] This result holds for almost all commonly used distributions, but there exist situations where it may not hold. The exact statistical conditions are provided in the Appendix.

We compare team and hybrid processes as opposed to team and purely individual processes. Organizations are by definition comprised of multiple individuals. In order to realize organizational objectives, at some point the efforts of individuals must be coordinated. Furthermore, we are interested in comparing organizational structures comprised of the same level of resources. Thus, we compare a team process (in our case comprised of four individuals) and the same number of individuals organized in a hybrid structure in which they first work individually and then spend a smaller amount of time together coordinating their activities. The hybrid process has a much higher component of individual work in comparison to the team process; thus phenomena that arise in individual settings are more likely to arise also in the hybrid process.

A vast body of research has examined the differences between team and individual idea generation. In a comprehensive series of studies, Diehl and Stroebe (1987), Diehl and Stroebe (1991), and Stroebe and Diehl (1994), identified that team brainstorming leads to *production blocking* (the inability to articulate ideas when others in the team are speaking), *evaluation apprehension* leading to censoring of potentially good ideas, and *free riding* (i.e., collective performance measures impeding individual incentives to perform). Further, they demonstrate that production blocking largely leads to impeding the *number* of ideas generated. In our study we compare a team process, in which individuals work collectively and a hybrid process, in which individuals work by themselves for a fraction of the time and collectively after that. Thus, we expect production blocking in the team process to lead to a smaller pool of ideas generated in the team process than in the hybrid process. Moreover, there is likely to be more evaluation apprehension in the team process; leading to fewer ideas generated in the team process than in the hybrid process. Finally, free riding limits the incentives to perform, leading to both fewer ideas and a lower average quality of ideas for the team process.

In a seminal ethnographic study, Sutton and Hargadon (1996) and Hargadon and Sutton (1997), the authors found that idea generation is largely a process of technology accumulation and brokering. On similar lines, we believe many ideas are generated out of access to user experiences, experiences with certain technologies, and application of creativity templates (Goldenberg, Lehmann and Mazursky (2001)). The success of such a process of employing previous experiences as creative stimuli is contingent on access and retrieval of these experiences. In a team setting, the participants have access not only to their own experiences as in an individual setting, but they also have partial access to the experiences of others via intergroup communication. This should lead to more creative stimuli which, in turn, should lead to more building up on previously expressed ideas. This increased *interactive build-up* in teams should lead to a larger pool of ideas, and may lead to superior quality of ideas and lower variance in quality of ideas, because built-up ideas may be similar in content and consequently also similar in quality.

Collaborative processes like the team process have previously been found to lead to consensus building and convergence (Sutton and Hargadon (1996), Fleming and Singh (2007)). In our context, we expect this consensus building or *collaborative convergence* to lead to expression of increasingly similar ideas that have similar quality, thus limiting variance in teams.

However, team ideation also involves a larger degree of combination and cross-fertilization of thoughts from disparate individuals with different skill sets. Such ideas derived from the interactive combinations of diverse knowledge components have higher uncertainty in the compatibility of the components brought together (since they come from disparate individuals) (Fleming (2001), Fleming and Sorenson (2001), Taylor and Greve (2006)). We believe this effect of lack of *component compatibility* creates more potential for both breakdown and collaborative success in teams than in individual idea generation, which leads to both very good and very bad ideas. Consequently, we would expect this effect to increase the variance observed in the quality of ideas generated in teams.

Next, we examine how all the above mentioned effects are likely to come together to influence the statistics of the pool of ideas generated.

Average Quality of Ideas: Free-riding in teams will lead to lower incentives to generate great ideas leading to worse average quality of ideas. On the other hand, the access to more creative stimuli in teams can potentially allow for more build-up on existing ideas which may lead to the creation of better ideas. On balance, the net effect will depend on the relative magnitudes of the two phenomena. Further, previous work on brainstorming has not found any consistent effects on average quality (see Diehl and Stroebe (1987)). Consequently, we cannot construct a hypothesis a priori from the literature on the net effect of the organizational process on the average quality of ideas generated. As a result, we pose a null hypothesis, which we can be tested with our experiment.

Hypothesis 1: The average quality of ideas generated from the team and hybrid processes is the same.

Number of Distinct Ideas Generated: Free riding, evaluation apprehension, and production blocking all suggest that teams will be able to generate fewer ideas. On the other hand, access to more creative stimuli and disparate knowledge components in teams can lead to the possibility of more combinations that lead to more distinct ideas. Again, the net effect will depend on the relative magnitudes of these effects. Previous research has found that production blocking is a very strong phenomenon and generally its effects far outweigh other phenomena (Diehl and Stroebe (1987)). In line with these observations, we hypothesize that the detrimental effects of production blocking, free-riding, and evaluation apprehension in teams will outweigh any benefits from more possibility of building up.

Hypothesis 2: The number of distinct ideas generated (per person per unit time) in the hybrid process is higher than the number of distinct ideas generated in the team process.

Variance in Quality of Ideas: The effect of collaborative convergence in teams and interactive build-up work to make the quality of ideas more similar, whereas the increased risks of knowledge component incompatibility lead to higher quality variance. The net effect of these phenomena will depend on their relative magnitudes. To the best of our knowledge, previous research does not provide any strong prescriptions on this, so we pose the null hypothesis:

Hypothesis 3: The variance in quality of ideas in the team and hybrid processes will be the same.

Build-Up of Ideas in Teams: We have argued that teams are more likely to build on previously mentioned ideas. Further, we argued that this build-up has a positive effect on quality and will tend to increase the number of ideas generated. Since our experimental set-up allows us to measure the extent to which a group builds on previous ideas, we can test the indirect effect of choice of organizational process on the quality, variance, and number of ideas. Note that these effects are indirect, because for example, the choice of organizational form may directly affect idea quantity but may also have an effect through its role in contributing to build-up. These effects are reflected in these three related hypotheses.

Hypothesis 4a: Teams generate a higher fraction of ideas that build on previous ideas than do hybrid groups.

Hypothesis 4b: Ideas that build on previous ideas are of higher average quality.

Hypothesis 4c: Building on previous ideas increases the productivity of the group.

3.2 The Idea Selection Step

In the idea selection step, the group evaluates and selects the most promising ideas from those originally generated. Since an objective measure of quality is typically not possible; organizational units usually build a subjective estimate of the future potential of each idea and use that to construct relative preferences. These estimates may or may not correlate well with the "true" quality of an idea.[3] A process that provides a more accurate measure of the relatively quality of different ideas on average should lead to the selection of higher quality ideas. As an extreme example consider two organizational processes– one that can perfectly discern the true quality of the ideas, and one that has no ability to distinguish between ideas of different quality. When presented with identical pools of ideas, the first process will select the true best subset of ideas. The second process on the other hand will select a random subset from the original pool. On average, the quality of the random subset will be inferior to the quality of the true best subset of ideas. For an organization interested in the quality of the best identified ideas, the *fidelity of the evaluation process* it employs is thus crucial.

From a statistical perspective we know that a process that has access to more independent, unbiased estimates of quality will be able to construct more accurate estimates of quality. There are two potential sources of bias and interdependence in the idea generation and selection process. First, if the same unit that created the idea is also asked to evaluate the idea, this unit may be biased in favor of its own ideas.

[3] The notion of "true quality" is challenging and several conceptual frameworks for true quality are possible. Because the value that is eventually realized from an idea is uncertain, one way to think about true quality is as the expected net present value of the idea if pursued in a value-maximizing fashion by the organization. This notion of value could in theory be generalized to accommodate non-financial value outside of commercial settings.

Furthermore, ideas that for one reason or another garnered discussion time in the creation phase are made salient and therefore most likely to be perceived as high quality by the team members. These sources of bias are more prevalent in the team process than in the hybrid process. This is because in the hybrid process, the majority of ideas are likely to have been created during the individual phase and then evaluated by others in the group phase, reflecting independence between creators and evaluators.

A second source of interdependence arises among group members in a team setting. Previous research has shown that team members affect one another's perceptions, judgments and opinions (Gibson (2001), Stasser and Davis (1981), Zander and Medow (1963)). Detailed observation of the team cognitive processes has found that often "high-status" members dominate the discussion (Bandura (1997), Bartunek (1984), Davis, Bray and Holt (1977), Gibson (2001), Laughlin and Shippy (2006)). Because of these effects, we believe that the aggregation of information in teams will reflect interdependence among group members, and thus will not result in estimates of quality that are as good as those of the hybrid process.

Hypothesis 5: The hybrid process will be more accurate in evaluating the generated ideas than the team process.

3.3 The Selected Best Ideas

In the two preceding sections, we developed theory for how the idea generation step and the idea selection step are influenced by the choice of organizational process. Many different effects influence each of the two steps. The phenomena that influence idea generation and those that influence idea selection come together to drive the quality of the best idea. The net effect of these multiple competing phenomena depends largely on their magnitudes and interactions. Since Hypotheses 2 and 5 favor the hybrid process while Hypothesis 4 favors the team process, at this point we are unable to state a hypothesis capturing the overall (net) effect. Instead, we again pose the null hypothesis:

Hypothesis 6: Team and hybrid processes are equally effective in generating and selecting a set of best ideas.

4 Experimental Design

To compare the effectiveness of teams and hybrid structures for creative problem solving, we ran an experiment that allowed us to compare the treatments with respect to their impact on the average quality of ideas generated, on the number of ideas generated (productivity), on the variance in quality, on the ability to discern quality, on the extent of interactive build-up, on the quality of the best generated ideas, and on the quality of the best selected ideas. We employ a within-subjects design for this study. In such a design, each subject generates ideas under *both* the treatments– team and hybrid. Such a design helps us control for any differences in individual ability, team composition, and team dynamics. Further, one property of interest, within-team variance in idea quality, needs to be separated from across-team quality variance. This is most effectively done in a within-subjects design. Figure 2 illustrates the experiment design.

The experiment was conducted in two phases: (1) an *idea generation and self-evaluation* phase where the subjects created and developed a consensus ranking of the best ideas (self evaluation), and (2) a completely separate *independent evaluation phase* where judges rated the quality of ideas and coded the content of ideas.

4.1 Idea Generation and Self-Evaluation Phase

Subjects: Subjects for the experiment were recruited from students in an upper-level product design elective course at the University of Pennsylvania. All subjects had participated in multiple brainstorming and idea generation exercises prior to the experiment and had received training in idea generation techniques. The 44 subjects came from a wide variety of majors, with a majority in engineering and business. Most subjects were juniors, seniors, or masters-degree candidates. All experiments were conducted after obtaining prior approval from the human subjects committee at the university and participation in the exercise was voluntary and had no bearing on performance in the course. The subjects were informed that this was as an experiment to understand the idea generation process. Since extrinsic incentives are known to limit creative behavior (Amabile (1996)), no explicit incentives or compensation were provided for participation or performance in the experiment.

Treatments: In the team idea generation process, subjects were divided randomly into teams of four. Each team was given 30 minutes to complete an idea generation challenge. The subjects were asked to record each idea on a separate sheet of paper. A pre-stapled and pre-ordered bundle of sheets was provided each team. The sheets included an area for notes related to the idea and a designated area to record a title and a 50-word description. At the end of 30 minutes, the subjects were given an additional 5 minutes and instructed to develop a consensus-based selection and ranking of the best 5 ideas generated by their team.

In the hybrid process, subjects were asked to work individually on an idea generation challenge for 10 minutes. At the end of 10 minutes, the individuals were asked to rank their own ideas. The subjects were then divided randomly into groups of 4 and given a further 20 minutes to share and discuss their ideas from the first phase and to develop new ideas. All ideas, from both the individual and group portion of the process, were recorded on sheets as described for the team process. At the end of the group phase of the hybrid idea generation process, subjects were given an additional 5 minutes and instructed to develop a consensus-based selection and ranking of the best 5 ideas generated by their group, including those generated as individuals.

Experiment: Participants were divided into two clusters– one cluster was administered the hybrid treatment first followed by the team treatment and the other was administered the team treatment first followed by the hybrid treatment. For each of the two clusters, half the subjects were given Challenge 1 for the first treatment followed by Challenge 2 for the second treatment, the other half were given Challenge 2 for the first treatment and Challenge 1 for the second treatment. The idea generation exercises are described below. This setup allowed us to control for effects arising out of the order of treatments, the order of the challenges, and/or related to interactions between the treatments and the challenges.

> **Challenge 1:** You have been retained by a manufacturer of sports and fitness products to identify new product concepts for the student market. The manufacturer is interested in any product that might be sold to students in a sporting goods retailer (e.g., City Sports, Bike Line, EMS). The manufacturer is particularly interested in products likely to be

> appealing to students. These products might be solutions to unmet needs or improved solutions to existing needs.
>
> **Challenge 2:** You have been retained by a manufacturer of dorm and apartment products to identify new product concepts for the student market. The manufacturer is interested in any product that might be sold to students in a home-products retailer (e.g., IKEA, Bed Bath and Beyond, Pottery Barn). The manufacturer is particularly interested in products likely to be appealing to students. These products might be solutions to unmet needs or improved solutions to existing needs.

A total of 443 ideas were generated and evaluated by the 44 subjects. A sample of ideas generated is provided in the Appendix.

4.2 Independent Evaluation Phase

Because an accurate measurement of idea quality is essential to the testing of our theory, we employed two measurement methods. We believe that these methods go well beyond the accuracy of measurement used in prior studies.

Business value of product idea: First, we measured the utility of the ideas to a commercial organization that could develop and sell the products. To assess this value, we assembled a panel of 41 MBA students, completely distinct from subjects involved with the first phase of the experiment, who had all received formal training in the valuation of new products through a series of graduate classes. This panel was asked to assess the business value of the generated product ideas using a scale from 1 (lowest value) to 10 (highest value). The ideas were presented independently to the panelists in a random order. Each panelist rated between 206 and 237 different ideas. Each idea was rated by at least 20 different members of the panel. To verify the reliability of these ratings, we follow the method prescribed by Gwet (2002). We constructed Kappa (8.99, 2.92) and AC1 (13.38, 7.59) statistics for each of the two idea domains. All statistics suggest very high levels overall reliability in classification of ideas on our 10 point scale.

Probability of Purchase: We also evaluated the product ideas from the perspective of potential consumers. For this exercise we enrolled 88 subjects who were representative of the target market for the product ideas generated. The two challenges focused on products for college students, and consequently we enrolled college students for this purchase-intent survey. The participants in the survey were provided descriptions of the product ideas and were asked to assess their likelihood of purchasing the products on a 10 point scale. The product descriptions were provided in a randomized order and each survey participant saw between 200 and 245 different ideas. Each idea was rated by at least 44 different potential customers following standard market research techniques on measuring purchase intent (cf. Ulrich and Eppinger (2007) and Jamieson and Bass (1989)). To verify the reliability of the ratings, we again follow the method prescribed by Gwet (2002). We constructed Kappa (11.45, 9.93) and AC1 (8.92, 11.627) statistics for each of the two idea domains. All statistics suggest very high levels of overall levels of reliability in classification of ideas on our 10 point scale.

Finally, previous research has characterized the quality of new products as multi-dimensional, including the dimensions of attractiveness and feasibility. We also created a multi-dimensional quality scheme composed of five different metrics: *technical feasibility* (to what extent is the proposed product feasible to develop at a reasonable price with existing technology), *novelty* (originality of the idea with respect to the unmet need and proposed solution), *specificity* (the extent to which the idea included a proposed solution), *demand* (reflecting market size and attractiveness), and *overall value*. To rate ideas on these dimensions, we recruited a team of two graduate students specializing in new product development and instructed them to rate each idea with respect to these dimensions on 10 point scale. We discarded all ratings for which the two raters disagreed by more than 2 points. Looking at the remaining ratings, we found that the five dimensions were highly correlated. Factor analysis suggested using only one composite factor for the five metrics. Further, each of the metrics was highly correlated with estimates of business value and probability of purchase which we constructed using larger panels. In light of this correlation and the apparent lack of independent underlying dimensions in the expert judgments, we will present our results using the business value and purchase probabilities from the two large panels of judges.

4.3 Measuring the Build-Up of Ideas

A key explanatory variable in our theory is the progressive build-up of ideas. To measure this build-up, we hired three independent judges to code the substance of ideas on different dimensions. Ideas generated in Challenge 1, sporting goods, were categorized along the following three dimensions: the type of product, the principal sporting activity associated with the product and the key benefit proposition of the proposed product. The coders were provided with a set of exhaustive and mutually exclusive potential categorizations for each of the three dimensions. These categories were developed by examining product classifications by the online retailers Amazon, Wal-Mart, and Buy.com. Unrepresented categories in the data were eliminated. As an example, the product idea “cleated shoe covers – a protection for shows with cleats, to enable walking on hard surfaces without damaging the cleats”, was categorized by our coders as footwear (type of product), field sports (principal sporting activity) and convenience (key benefit proposition). The full list of categories for each of the three dimensions is provided in the Appendix.

Ideas generated in Challenge 2, products for a student residence, were categorized in a similar manner. The corresponding dimensions were product category, the typical room or location of that product and the key benefit. The full list of categorizations for ideas generated for Challenge 2 is in the Appendix.

To construct our build-up metric, we compare the classification of two consecutively generated ideas. For example, if the idea shares all three dimensions with the idea that was generated immediately before this idea, it earns a build-up score of 3. More generally, the build-up score is the number of dimensions that an idea shares with the idea generated immediately previously. We average this build-up score across the three independent judges.

5 Effect of the Idea Generation Process on Mean Quality, Number of Ideas Generated, and Variance of Quality

In this section, we report the results concerning Hypotheses 1-3. All hypotheses related to idea quality are tested using both business value and purchase intent as measures of quality. Unless stated otherwise, we use an ANOVA analysis of the judges’ ratings given each idea. That is, each rating of an idea provided by an independent judge is the dependent variable for a separate observation. The explanatory variable is the treatment (team vs. hybrid). We include controls for the four-person group of individuals generating the ideas (the "creator") and the rater who provided the rating. This is because there are substantial differences in ability across the groups, and because there are systemic differences in how the scales were used by different raters. We considered the rater and creator effects as both fixed effects and random effects. Our results are nearly identical in either case. Further, a Hausman test verifies the appropriateness of the use of the random effects estimators.[4]

5.1 Effect of Idea Generation Process on the Mean Quality

Table 2, row 5.1, shows the results for the mean quality for the two different treatments. We evaluate and test the statistical significance of the difference in quality and are able to reject Hypothesis 1, finding that the *hybrid process generates ideas of better average quality*. The quality advantage of the hybrid treatment is 0.25 units of business value and 0.35 units of purchase intent (significant at the 0.01% level for both business value and purchase intent). Although the magnitude of this difference may not appear large relative to the 10-point scale, a difference this large can roughly translate to about 30 points in percentile ranking (after controlling for fixed effects), in other words, this can be the difference between the 1st and the 30th idea in a pool of 100 ideas.

5.2 Effect of Idea Generation Process on Productivity (Number of Ideas Generated)

Table 2, row 5.2, illustrates the results of an ANOVA analysis of the productivity, or the number of ideas generated in the two treatments, given the same number of people working for the same amount of time. The value shown is the number of ideas generated by the four-person group in 30 minutes. We control for the effects of the sets of individuals generating ideas and consider two alternate specifications, one with the creators as a random effect and a repeated measures analysis. Our results are almost identical in the different specifications. We find that the productivity is very different across different treatments; the

[4] The Hausman test compares the estimates from the more efficient random effects model against the less efficient but consistent fixed effects model to make sure that the more efficient random effects model also gives consistent results.

hybrid process generates about three times more ideas than the team process (significant at the 0.01% level). This result supports Hypothesis 2 and the existing literature. To the best of our knowledge we are the first to verify these results statistically in a within-subjects design that controls for individual effects.

5.3 Effect of Idea Generation Process on the Within-Group Variance in Idea Quality

As argued in Section 3.1, the variance in quality of ideas generated by each group under the two different treatments influences the quality of the best idea. Note that this is not the variance in the quality ratings of the ideas across treatments or across groups but the variance in the quality of the ideas *within a particular group*. We define this variance measure as the squared difference of the rating received by an idea and the average rating received by all ideas generated by the group in the specific treatment. We then conduct an ANOVA for this variable. The results are reported in Table 2, row 5.3. We do not find any evidence for a difference between the team process and hybrid process as far as the variance of idea quality is concerned. Thus, we are not able to reject Hypothesis 3.

6 Net Effect of Idea Generation Process on the Best Ideas (Extreme Values)

In the preceding section, we examined how the team process and the hybrid process of idea generation differ along the four variables that determine the quality of the best idea in the context of our theoretical framework (Figure 1). In this section, we will examine how these properties come together to influence the quality of the best generated ideas and the best selected ideas.

6.1 Quality of the Best Generated Idea

Given our results that relative to the team process the hybrid idea generation process has higher mean quality, higher productivity, and equivalent variance, we expect that the quality of the best generated ideas to be higher for the hybrid process.

Hypothesis 7: The quality of the best generated ideas will be higher in the hybrid process.

To test this hypothesis we conduct an ANOVA analysis of the ratings received by the top 5 ideas generated by each group. Table 2, row 6.1, shows the results from the comparison of the average quality of top 5 ideas in different treatments. We also test alternate versions of this hypothesis, with the top 3, 4 and 6 ideas. In each of these cases our results provide similar support. As before, we include controls for the group of individuals generating the ideas, the rater who provided the rating, and the challenge to which the idea is addressed.

The ANOVA shows that the team and hybrid process are different in the quality of the top 5 ideas. In particular, we evaluate and test the statistical significance of this difference and find that, as predicted in Hypothesis 7, the top 5 ideas from the *hybrid process are of better quality* than those from the team process. Interestingly, the difference between the team and hybrid in terms of the quality of *best* ideas is much higher than the difference in *mean* quality of ideas. This follows from our previous observations related to productivity and variance of quality. Further, it illustrates that in an innovation setting, examining only mean quality as opposed to the quality of the best ideas is likely to underestimate the benefits of the hybrid approach.

6.2 Effect of Idea Generation Process on Ability to Discern Quality

We measure the ability to discern quality as the rank correlation between the preference ordering implied by the independent judges' ratings and the self evaluation by the idea generating group. As with all previous results, we provide this analysis for both business value ratings and the purchase intent ratings. The results are provided in Table 3. Note that the absolute value of the correlation for either team or hybrid is relatively low, in the best case less than 0.2. This suggests that irrespective of the process, team or hybrid, the ability of idea generators to evaluate their own ideas is extremely limited, and is perhaps compromised by their involvement in the idea generation step. Secondly, the hybrid process has a significantly higher ability than the team process, supporting Hypothesis 5. In further analysis, we compared the self evaluation provided in the individual phase of the hybrid treatment to the independent judges' quality ratings, and find that these individual ratings are better predictors of "true quality" than are either of the group evaluations, lending further support to the idea that some aspect of the group interaction leads to poor assessments of quality.

6.3 The Quality of the Best Selected Ideas

The creative problem solving process includes both idea generation and idea selection. In this section, we will include the impact of idea selection in our analysis. To do so, we compare the quality of the top 5 *selected* ideas between the hybrid and team organizational processes. To test this hypothesis we conduct an ANOVA on the independently determined quality ratings for the top 5 selected ideas. Table 2, row 6.3, shows us the results from the comparison of the average quality of top 5 selected ideas in different treatments. For the purchase-intent quality metric we can reject Hypothesis 6, concluding that the hybrid process results in higher quality for the best 5 selected ideas. For the business-value quality metric, we are not able to reject the hypothesis that both treatments result in top 5 ideas of equal quality. These results suggest that the hybrid process may generate better ideas, but that due to the noisy selection process, its relative advantage is much diminished, to the point of becoming statistically insignificant for one of our quality metrics.

7 Analyzing the Mechanisms of Action: Building up on Ideas

The results of the previous sections show that the hybrid process generates better ideas. Thus, the interactive build-up effect theorized for teams must be weak, at least when compared to the other effects in our theoretical framework. Our experimental design allows us to measure the extent of build-up at the idea level. In particular, recall that we coded the content of all ideas and computed the content similarities between consecutive ideas, which gives us a metric of the extent of build-up for these ideas.

In this section, we first test if individuals working in teams are more likely to build up on ideas than individuals working in the group phase of the hybrid process (Hypothesis 4a). Next, we will investigate the impact of this build-up on the variables that drive mean idea quality (Hypothesis 4b) and productivity (Hypothesis 4c).

7.1 More Build-Up in Teams?

The existing literature has argued that teams are more likely to build up on ideas. Recall that the build-up score is a measure of the extent to which an idea is similar to the previous idea. Table 2, row 7.1, shows the results from an ANOVA of the build-up scores of ideas. The results support Hypothesis 4a and the observation in the literature that ideas generated in teams are more likely to build on previous ideas.

7.2 Impact of build-up on Mean Quality of Ideas Generated

To investigate the impact of build-up on mean quality, we cannot conduct a direct regression (nor ANOVA) of quality on build-up. Such an approach would lead to incorrect estimates as both quality and build-up are influenced by an omitted variable in this regression, the choice of organizational process. In other words, the error term in such a direct regression will include the effect of the process and this would be correlated with the dependent variable. Thus, to test this effect we propose a two-stage least-squares procedure. The estimated equations, the proposed path model and the standardized results from this model are illustrated in Figure 3.

The results of our path analysis confirm the previously observed direct effect of choice of organizational process on the quality and the extent of build-up. However, we find no support for the often-cited effect of build-up on improving quality of ideas. Thus, Hypothesis 4b is not supported. In fact, in one of our models, we find the *reverse* effect: due to increased build-up, we observe that the mean idea quality actually decreases. This suggests that while teams indeed build on each other’s ideas, this does not improve the quality of the ideas.

7.3 Impact of Build-Up on Number of Ideas Generated

Next, we analyze the impact of build-up on the number of ideas generated. We hypothesized that the interactive nature that leads to more build-up should expand the number of opportunities that a group identifies (Hypothesis 4c). To test this effect, we compute the average build-up in a group (following the team or hybrid process) and examine its impact on the number of ideas generated by the group. We follow the same empirical methodology as in the previous section. The estimated equations, the proposed path model and the standardized results from this model are illustrated in Figure 4.

Again, while there is more build-up in groups that followed the team process, this build-up has no impact on increasing the number of ideas generated. This again demonstrates that the beneficial consequences of build-up may have been over-estimated in the prior literature. One explanation for this is the competing effect of production blocking is so strong that it completely dominates the productivity gain from build-up.

8 Conclusions and Managerial Implications

In this study, we compare the effectiveness of two processes for a group of individuals solving problems that require creative idea generation followed by selection. First, the group of individuals can work as a team. Alternately, in a hybrid process, the group works individually for some fraction of the time followed by group work. We find strong support that the best ideas generated by a hybrid process are better than the best ideas generated by a group process. This result is driven by the fact that the hybrid process generates about three times as many ideas per unit of time and that these ideas are significantly higher quality on average. The hybrid process is also better at identifying the best ideas, however, we find that both approaches do poorly in absolute terms in selecting the best ideas. Our findings shed light on one of the longstanding arguments for team process, the benefits of interactive build-up. We show that the suggested advantage of team-based brainstorming is not supported by experimental evidence. On average, ideas that build on other ideas are not statistically better than any random idea. This has significant managerial implications: if the interactive build-up is not helping create better ideas, an organization might be better off relying on the asynchronous idea generation of individuals using, for example, web-based idea management systems, as this would ease other organizational constraints such as conflicting schedules of team members and travel requirements.

As with any experimental study, we have to caution the reader about generalizing our results. Our results on the quality of the best ideas depend not just on the directional comparisons between the two processes, but also on the magnitude of these differences. While our experiment was set up to closely match problems in real-world settings, the subjects’ limited time, resources, and prior exposure to the problem solving context limit our ability to perfectly mimic a real situation. Furthermore, while the subjects were trained in ideation techniques and knew each other somewhat, they were not placed in teams that had developed a great deal of collective experience.

In all our results, we found that differences in performance *across individuals* are large and highly significant. The large performance differences also suggest an interesting opportunity for future research. It would be interesting to examine if these differences are persistent. If they are, an optimal process may be to first screen the pool of individuals for the highest performers and then employ only them in subsequent idea generation efforts. However the dynamics of the interaction between these high-ability individuals may differ significantly from the existing evidence and need to be monitored in further experiments.

References

AMABILE, T. M. (1996): *Creativity in Context*. Boulder, CO: Westview Press.

BANDURA, A. (1997): "Self Efficacy," NJ: Prentice Hall.

BARTUNEK, J. (1984): "Changing Interpretive Schemes and Organizational Restructuring: The Example of a Religious Order," 355-372.

COLES, S. (2001): *An Introduction to Statistical Modeling of Extreme Values*. London: Springer Verlag.

DAHAN, E., and H. MENDELSON (2001): "An Extreme Value Model of Concept Testing," *Management Science*, 47, 102-116.

DAVIS, J. H., R. M. BRAY, and R. W. HOLT (1977): "The Empirical Study of Decision Processes in Juries: A Critical Review."

DIEHL, M., and W. STROEBE (1987): "Productivity Loss in Idea-Generating Groups: Toward the Solution of a Riddle," *Journal of Personality and Social Psychology*, 53, 497-509.

— (1991): "Productivity Loss in Idea-Generating Groups - Tracking Down the Blocking Effect," *Journal of Personality and Social Psychology*, 61, 392-403.

FLEMING, L. (2001): "Recombinant Uncertainty in Technological Search," *Management Science*, 47, 117-132.

FLEMING, L., and J. SINGH (2007): "The Lone Inventor as the Source of Technological Breakthroughs: Myth or Reality?," Harvard Business School.

FLEMING, L., and O. SORENSON (2001): "Technology as a Complex Adaptive System: Evidence from Patent Data," *Research Policy*, 30, 1019-1039.

GIBSON, C. B. (2001): "From Knowledge Accumulation to Accommodation: Cycles of Collective Cognition in Work Groups," *Journal of Organizational Behavior*, 22, 121-134.

GOLDENBERG, J., D. R. LEHMANN, and D. MAZURSKY (2001): "The Idea Itself and the Circumstances of Its Emergence as Predictors of New Product Success," *Management Science*, 47, 69-84.

GWET, K. (2002): *Handbook of Inter-Rater Reliability*. STATAXIS Publishing Company.

HARGADON, A., and R. I. SUTTON (1997): "Technology Brokering and Innovation in a Product Development Firm," *Administrative Science Quarterly*, 42, 716-749.

JAMIESON, L., and F. BASS (1989): "Adjusting Stated Purchase Intentions Measures to Predict Trial Purchase of New Products," *J Market Res*, 26, 336-345.

KAVADIAS, S., and S. SOMMER (2007): "The Effects of Problem Structure and Team Expertise on Brainstorming Effectiveness," Georgia Institute of Technology.

LAUGHLIN, P. R., and T. A. SHIPPY (2006): "Collective Induction," Psychology Pr.

MULLEN, B., C. JOHNSON, and E. SALAS (1991): "Productivity Loss in Brainstorming Groups: A Meta-Analytic Integration," *Basic and Applied Social Psychology*, 12, 3-24.

OSBORNE, A. F. (1957): *Applied Imagination*. New York: Charles Scribner's Sons.

PAULUS, P. B., V. BROWN, and A. H. ORTEGA (1996): "Group Creativity," in *Social Creativity in Organizations*, ed. by R. E. Pursuer, and A. Montuori. Creskill, NJ: Hampton.

ROBBINS, S. P., and T. A. JUDGE (2006): *Organizational Behavior*. Upper Saddle river, NJ: Prentice Hall.

STASSER, G., and J. H. DAVIS (1981): "Group Decision Making and Social Influence: A Social Interaction Sequence Model," 523-551.

STROEBE, W., and M. DIEHL (1994): "Why Are Groups Less Effective Than Their Members: On Productivity Losses in Idea Generation Groups," *European Review of Social Psychology*, 5, 271-303.

SUTTON, R. I., and A. HARGADON (1996): "Brainstorming Groups in Context: Effectiveness in a Product Design Firm," *Administrative Science Quarterly*, 41, 685-718.

TAYLOR, A., and H. R. GREVE (2006): "Superman or the Fantastic Four? Knowledge Combination and Experience in Innovative Teams," *The Academy of Management Journal*, 49, 723-740.

TERWIESCH, C., and C. H. LOCH (2004): "Collaborative Prototyping and the Pricing of Custom-Designed Products," Institute for Operations Research and the Management Sciences, 145-158.

TERWIESCH, C., and K. T. ULRICH (2009): *Innovation Tournaments: Creating and Selecting Exceptional Opportunities*. Harvard Business School Press.

ULRICH, K. T., and S. EPPINGER (2007): *Product Design and Development*. McGraw-Hill Higher Education.

ZANDER, A. W., and H. MEDOW (1963): "Individual and Group Aspiration," 89-105.

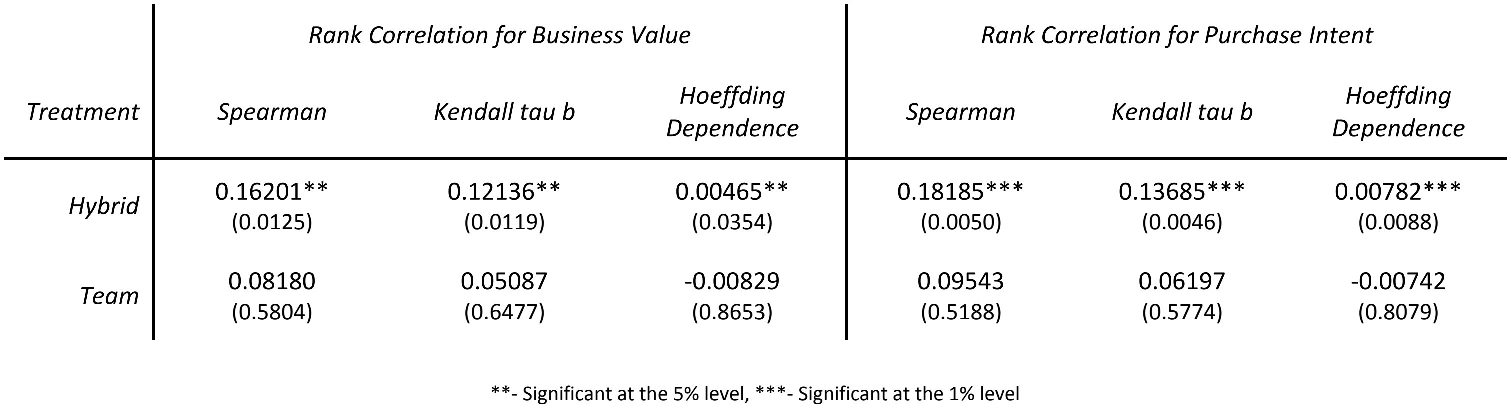
Research	Setting/Methodology	Measure of Idea Quality	Metrics	Results
Osborne (1957)				Introduced Brainstorming
Social psychology literature, summarized by Diehl & Stroebe (1987,1991, 1994)	Lab, Experimental	Rating by an assistant (Second assistant used for reliability) Rating by an expert	Mean quality & Productivity	Productivity: Individual > Teams Mean Quality: Equivocal Results No Reason to work in teams!
Sutton & Hargadon (1996,..)	Industry (IDEO), Observational			Contextual differences between lab and the real world
Taylor & Greve (2006)	Comic book industry, Empirical	Collector market value of a comic	Mean quality & Variance	Variance: Teams > Individuals Moderating effects of knowledge diversity, team experience, workloads, tenure, organizational resources
Fleming (2007)	Patent data, Empirical	No of patents, citations (use of patent)	Mean quality, Variance & Productivity	Mean: Teams > Individuals Variance: Individuals > Team
Kavadias & Sommer (2007)	Analytical			Depends on problem structure and team diversity (experience and knowledge)
Dahan & Mendelson (2001)	Analytical	Best idea (extreme value)	Extreme value of quality	
Girotra, Terwiesch & Ulrich	Lab (with trained subjects), Experimental	Ratings by a large number of peers using a web based interface	Mean quality, Variance, Productivity, Self-evaluation ability, Quality of *best* idea	[Reported in Sections 5 and 6]

Table 1: Summary of literature with comparison to this study.

Discussion Section	*Statistic Compared*	*N*	*F-Statistic for Team/Hybrid†*	*Least Square Mean Estimate for Hybrid‡*	*Least Square Mean Estimate for Team‡*	*Difference of Least Square Means: Hybrid-Team*
5.1	**Mean Quality&**					
	Business Value *(1-10 scale)*	8950	22.50***	4.79	4.52	0.265*** (4.74)
	Purchase Intent *(1-10 scale)*	18841	71.35***	4.93	4.58	0.349*** (8.45)
5.2	**Mean Productivity$** *(ideas per group per 30-min)*	22	26.23***	28.45	11.82	16.636*** (5.12)
5.3	**Within-Team Variance&**					
	Business Value	8950	2.34	6.42	6.63	-0.213 (-1.53)
	Purchase Intent	18841	2.41	8.23	8.06	0.169 (1.55)
6.1	**Quality of Top 5 Generated Ideas&**					
	Business Value	2157	69.55***	6.03	5.18	0.852*** (8.34)
	Purchase Intent	4535	151.14***	6.20	5.30	0.896*** (12.29)
6.3	**Quality of Top 5 Selected Ideas&**					
	Business Value	5720	2.95	4.63	4.77	-0.149 (-1.72)
	Purchase Intent	11841	24.91***	4.95	4.63	0.319*** (4.99)
7.1	**Degree of Build-up&**	7745	19.42***	2.20	2.41	-0.212*** (-4.41)

*** Significant at the <0.01% level. &: The unit of analysis is Idea-Rating. $: The unit of analysis is Organizational Unit. †: Results are reported from an ANOVA analysis with random effects for Raters and/or Creators. Identical results are obtained when raters and/or creators are introduced as fixed effects. ‡: Least Square means are the mean residuals after taking into account the other control variables.

Table 2: Results comparing team and hybrid treatments for each of dependent variables.



	Rank Correlation for Business Value			*Rank Correlation for Purchase Intent*		
Treatment	*Spearman*	*Kendall tau b*	*Hoeffding Dependence*	*Spearman*	*Kendall tau b*	*Hoeffding Dependence*
Hybrid	0.16201** (0.0125)	0.12136** (0.0119)	0.00465** (0.0354)	0.18185*** (0.0050)	0.13685*** (0.0046)	0.00782*** (0.0088)
Team	0.08180 (0.5804)	0.05087 (0.6477)	-0.00829 (0.8653)	0.09543 (0.5188)	0.06197 (0.5774)	-0.00742 (0.8079)

- Significant at the 5% level, *- Significant at the 1% level

Table 3: Rank correlation between self-assigned ranks and true ranks.

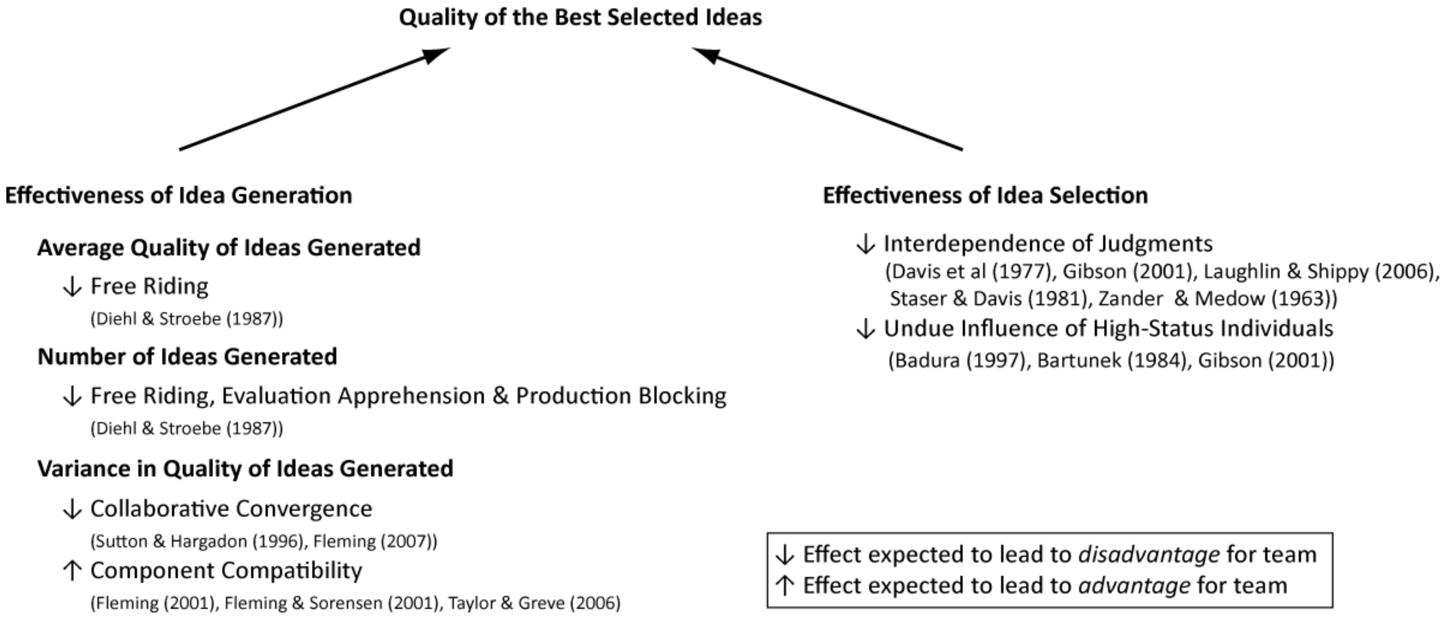

Quality of the Best Selected Ideas
Effectiveness of Idea Generation
Average Quality of Ideas Generated
↓ Free Riding
(Diehl & Stroebe (1987))
Number of Ideas Generated
↓ Free Riding, Evaluation Apprehension & Production Blocking
(Diehl & Stroebe (1987))
Variance in Quality of Ideas Generated
↓ Collaborative Convergence
(Sutton & Hargadon (1996), Fleming (2007))
↑ Component Compatibility
(Fleming (2001), Fleming & Sorensen (2001), Taylor & Greve (2006)
Effectiveness of Idea Selection
↓ Interdependence of Judgments
(Davis et al (1977), Gibson (2001), Laughlin & Shippy (2006), Staser & Davis (1981), Zander & Medow (1963))
↓ Undue Influence of High-Status Individuals
(Badura (1997), Bartunek (1984), Gibson (2001))
↓ Effect expected to lead to *disadvantage* for team
↑ Effect expected to lead to *advantage* for team

Figure 1: Model of creative problem solving process with hypothesized causal factors and links to the prior literature.

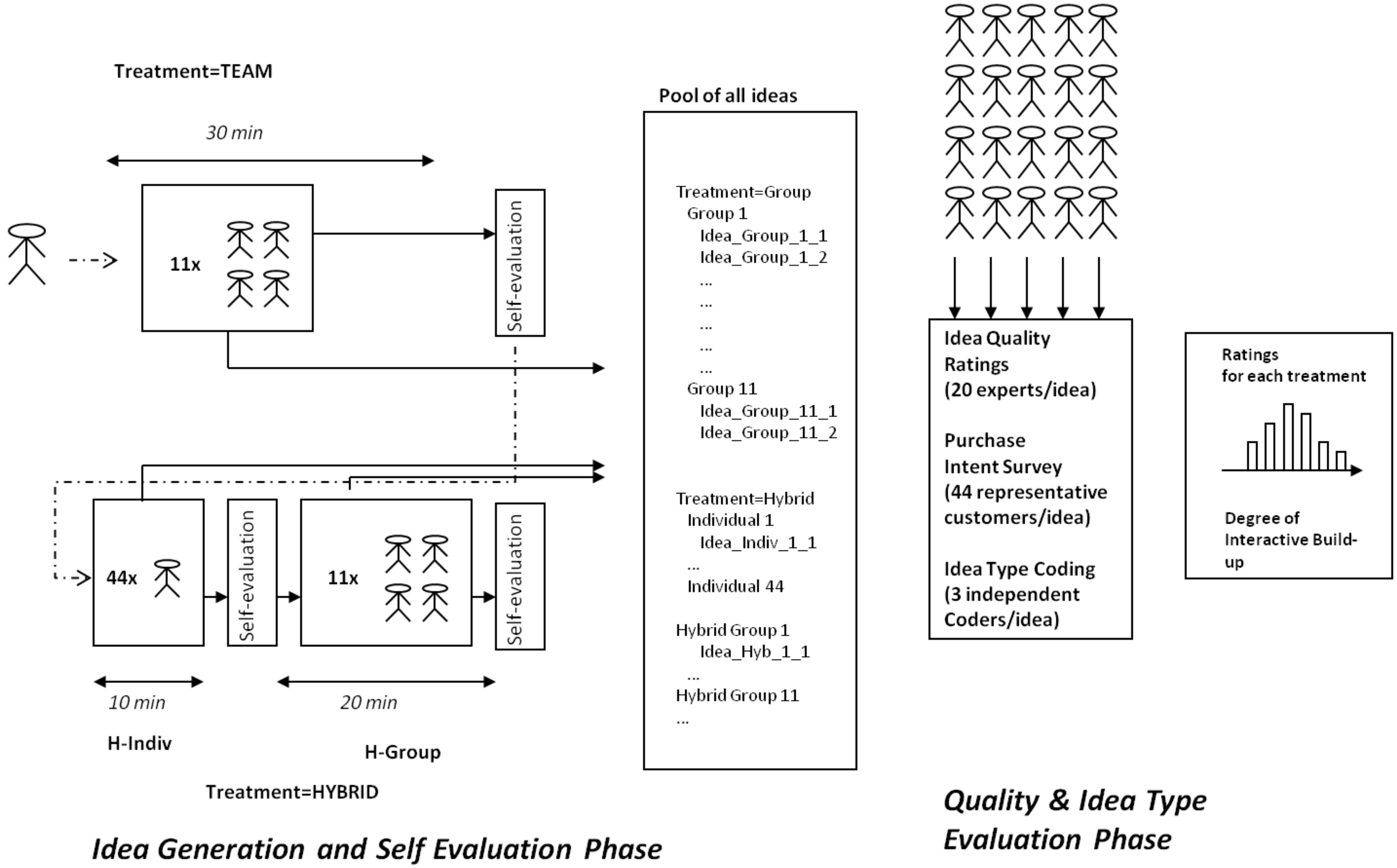
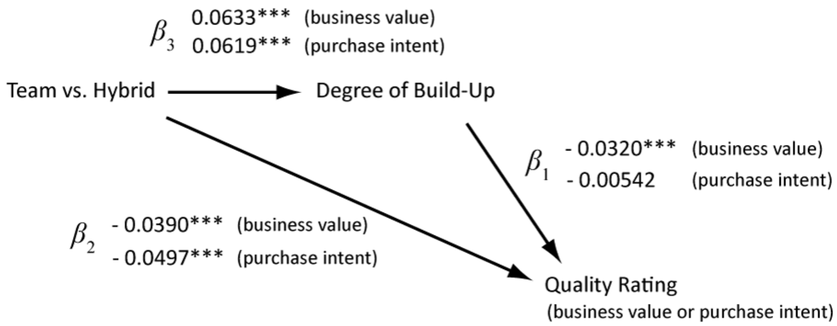


Figure 2: Experiment Design

$$\textbf{\textit{Build-Up}}_i = \alpha' + \beta_3\ \textbf{\textit{Team-v-Hybrid}}_i$$

$$\textbf{\textit{Quality-Rating}}_{ij} = \alpha + \beta_1\ \textbf{\textit{Build-Up}}_i + \beta_2\ \textbf{\textit{Team-v-Hybrid}}_i + \beta_4\ \textbf{\textit{Rater}}_j$$


β_3 0.0633*** (business value)
0.0619*** (purchase intent)
Team vs. Hybrid
Degree of Build-Up
β_1 - 0.0320*** (business value)
- 0.00542 (purchase intent)
β_2 - 0.0390*** (business value)
- 0.0497*** (purchase intent)
Quality Rating
(business value or purchase intent)

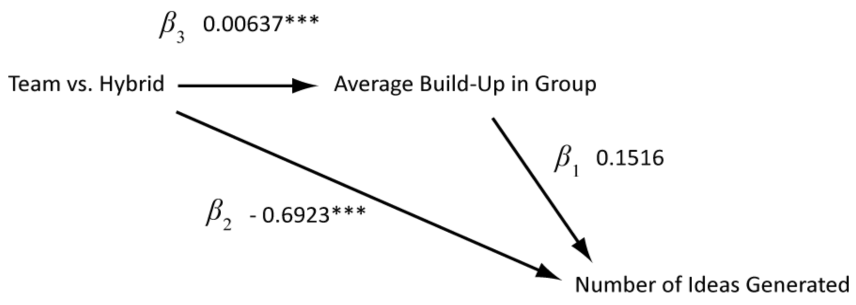
***- Significant at the 1% level.

Results are presented with standardized coefficients obtained from a MLE of the 2SLS model. The subscript *i* is an index for the idea and *j* indexes the rater.

Figure 3: Two-stage least-squares model and coefficient estimates for effect of Build-Up on idea quality (Business Value: N=7623, Purchase Intent N=16047).

$$\textit{Average-Build-Up-in-Group}_{kl} = \alpha' + \beta_3\ \textit{Team-v-Hybrid}_{kl}$$

$$\textit{N-Ideas}_{kl} = \alpha + \beta_1\ \textit{Average-Build-Up}_{kl} + \beta_2\ \textit{Team-v-Hybrid}_{kl}$$



***- Significant at the 1% level.

Results are presented with standardized coefficients obtained from a MLE of the 2SLS model. The subscript *k* is an index for the group and *l* is an index for the organizational process or treatment

Figure 4: Two-stage least-squares model and coefficient estimates for effect of Build-Up on the number of ideas generated (N=22).

This page is intentionally blank to separate the paper from its Appendix.

Appendix

1 Formal Statement of Theorems and Proofs from Section 3.1

Theorem 1(Effect of Number of Ideas): $\boldsymbol{E}[M_n] \leq \boldsymbol{E}[M_{n+1}]$

Proof: Note that the $\Pr[\mathrm{M_n} \leq \mathrm{z}] = \prod_{\mathrm{i=1}}^{\mathrm{n}} \Pr[\mathrm{X_i} \leq \mathrm{z}]$. Thus, the Cumulative Distribution Function of the distribution of M_n, $G(z)$ is $F^n(z)$. $\mathbf{E}[M_n] = \int_0^\infty zg(z)dz = \int_0^\infty \left(1 - G(z)\right)dz = \int_0^\infty \left(1 - F^n(z)\right)dz$. Since $F(z) \leq 1$, $F^{n+1} \leq F^n$ and $1 - F^{n+1} \geq 1 - F^n$. The result now follows. ■

Lemma 1: If the quality of ideas generated follows a Generalized Extreme Value Distribution (GEV) (Coles (2001)) with parameters (μ, σ, ξ) the quality of the best of n ideas also follows a Generalized Extreme Value distribution with parameters

$$\mu^{'} = \mu + \frac{\sigma}{\xi}\left(n^{\xi} - 1\right)$$
$$\sigma^{'} = \sigma n^{\xi}$$
$$\xi' = \xi$$

Proof: The result follows from substituting the cumulative distribution functions and reparameterizing. ■

A similar result has been shown by both Dahan and Mendelson (2001) and Kavadias and Sommer (2007). While Dahan and Mendelson (2001) work with the three different sub-families of the generalized extreme value distributions, we present our result within the unifying framework of the generalized extreme value distribution. Kavadias and Sommer (2007) present this result for the Gumbel Distribution. Also, note that the generalized extreme value distribution represents a fairly flexible family of distributions that can capture a wide variety of censored data. Since idea generation often involves some internal censoring by the ideator, this family is an ideal candidate for capturing idea quality. Further, from data collected under a variety of ideation settings in real organizations, we find this family to be a reasonable fit.

Theorem 2 (Effect of the mean of the idea quality distribution) Consider two ideation processes with GEV quality distributions with different means. All other central moments of the distributions are identical.

The processes generate the same number of ideas. The expected quality of the best idea from the ideation process with the higher mean is higher.

Proof: Since all moments besides the mean are identical for the two distributions, only the location parameter of the two quality distributions μ can be different say $\mu_1 > \mu_2$. From Lemma 1, the best idea from each of the ideation processes will also be distributed GEV, with all parameters identical except the location parameters $\mu_1^{'} > \mu_2^{'}$. The mean of GEV distribution increases in the location parameter and the result now follows. ■

This result shows that all else being equal, the quality of the best idea from a process with a higher average quality is higher.

Theorem 3 (Effect of the variance of the idea quality distribution): Consider two ideation processes with GEV quality distributions with different variance. All other central moments of the distributions are identical. The processes generate the same number of ideas. The expected quality of the best idea from the ideation process with the higher variance is better iff $\Gamma(1-\xi) > 0$

Proof: Consider two GEV distributions (μ_1, σ_1, ξ_1) and (μ_2, σ_2, ξ_2). The conditions on the central moments of the two distributions imply that $\xi_1 = \xi_2 = \xi$. $\sigma_1 \neq \sigma_2$; say $\sigma_1 > \sigma_2$ and $\mu_1 - \mu_2 = (\sigma_1 - \sigma_2)\frac{(1-\Gamma(1-\xi))}{\xi}$. From Lemma 1, the quality of the best idea from each of the ideation processes will also be distributed GEV, with parameters $\left(\mu_1 + \frac{\sigma_1}{\xi}\left(n^{\xi}-1\right), \sigma_1 n^{\xi}, \xi\right)$ and $\left(\mu_2 + \frac{\sigma_2}{\xi}\left(n^{\xi}-1\right), \sigma_2 n^{\xi}, \xi\right)$ and means $\mu_1 + \frac{\sigma_1}{\xi}(n^{\xi}\Gamma(1-\xi)-1)$ and $\mu_2 + \frac{\sigma_2}{\xi}(n^{\xi}\Gamma(1-\xi)-1)$, Γ is the gamma function. The result will hold if $\frac{(n^{\xi}-1)\Gamma(1-\xi)}{\xi} > 0$. Now note $n > 1 \Longrightarrow \frac{(n^{\xi}-1)}{\xi} > 0$. The result follows. ■

Corollary: Consider two ideation processes with Gumbel quality distributions with different variances. All other moments of the distributions are identical. The processes generate the same number of ideas. The expected quality of the best idea from the ideation process with the higher variance is better.

Proof: The Gumbel distribution belongs to the GEV family with $\xi \to 0$. The result follows from an application of the above theorem and assuming n > 1. ■

Theorem 4: a) (Coles (2001)) If there exist sequences of constants $\{a_n,\ b_n\}$ *such that*

$$\Pr\{M_n^* \leq z\} \to G(z)\ as\ n \to \infty$$

for a non-degenerate distribution function G, then G is a member of the GEV family

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

defined on $\{z: 1 + \xi(z-\mu)/\sigma > 0\}$, *where* $-\infty < \mu < \infty$, $\sigma > 0$ *and* $-\infty < \xi < \infty$.

b) Given $\{Z_1, Z_2, \ldots, Z_m\}$, *m observations of* M_n, *the parameters of* $G(z)$ *can be estimated as the argmax of the log-likelihood function*

$$l(\mu, \sigma, \xi) = -m\ log\ \sigma - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{m} log\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right] - \sum_{i=1}^{m}\left[1 + \xi\left(\frac{z_i - \mu}{\sigma}\right)\right]^{-1/\xi}$$

provided that $1 + \xi\left(\frac{z_i - \mu}{\sigma}\right) > 0$, *for i=1,...,m. As always with maximum likelihood estimation, the parameter estimates are asymptotically normally and approximate confidence intervals can be constructed using the observed information matrix.*[5]

Proof a) The result is well known and we refer the reader to Coles (2001) for an outline of the proof and to the references therein for a more technical version of the proof.

[5] A potential difficulty with the use of maximum likelihood methods for the GEV concerns the regularity conditions that are required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid. These conditions are not satisfied by the GEV model because the end-points of the GEV distribution are functions of the parameter values: $\mu - \sigma/\xi$ is an upper end point of the distribution when $\xi < 0$, and a lower end point when $\xi > 0$. Smith (1985) considers this problem in detail and find that for $\xi > -1$, the estimators are generally obtainable and often have the usual asymptotic properties.

b) Under the assumption that $\{Z_1, Z_2 \dots, Z_m\}$ are independent variables having the GEV distribution, the above log likelihood follows from simple computation and absorbing the constants within the estimated parameters in the usual way. ■

2 Subsample of Ideas Generated

Title	Descriptions	Mean Rating
Mouth guard Holder	A small, convenient, removable pocket that can be used to hold a mouth guard in between uses on the field.	4.1
Odor Reducing Trash Can	A trash can that reduces odor of garbage inside it.	6.5
Water Bottle with Filter System	A water bottle with a built-in filtration system.	5.9
Transforma-Racquet	An athletic racquet that can be adjusted to accommodate any racquet sport.	4.2
Waterproof Reading System	A system for reading in the shower.	3.2
Disposable Desktop Cover	This product is meant to be placed over a clean desktop. As clutter builds up, just fold up the cover and pull the draw string to trash the collected garbage.	3.5
Toilet Table	A foldable table that attaches to the toilet so you can read, eat, or do work while going to the bathroom.	3.8
Coffee Table with Built-in Remote	A coffee table that has a TV remote built into it so that you don't have to move far to change channels, but at the same time you don't have to search for a lost remote.	3.7
Ball Bag	A ball that functions as a bag until it is time to use it. When the ball is emptied, it then turns into a ball to use.	3.4
Motion Detection Light	A light that detects that someone is trying to turn it on. When it senses motion at close proximity to the senor, it will automatically turn on or off.	3.6
Hair Collecting Comb	A comb that collects stray hairs and makes them easy to dispose.	5.3
Chore Meter	A system that logs who did what chores at a certain time to establish who isn't carrying their load.	3.9
Noise Reduction Pad	A pad that is placed on the floor of a dorm room to reduce the level of noise heard by the room below. Designed for students that work out in their rooms.	5.5

3 Idea Categorization Scales

3.1 Challenge 1: Sports and Recreation

Ideas generated in challenge 1 (sports and fitness products) were classified along the dimensions of "Type of Product", "Principal Sporting Activity" and "Key Benefit Proposition" in the following categories:

Type of Product	*Principal Sporting Activity*	*Key Benefit Proposition*
Bag	Basketball	Convenience

Bottle	Bicycling	Hi-Tech
Clothing	Field Sports	Multipurpose
Gear and Equipment	Golf	Hygiene
Food and Drink	Gym / Strength / Fitness	Portability
Locks / Security	Tennis and Racquet Sports	Customization / Personalization
Music / Entertainment	Running	Weather protection
Footwear	Swimming	Health
Information Systems	Winter Sports	Style
Watch	Not specific to activity	Reminder
	Other sport/activity	Eco-friendly

3.2 Challenge 2: Dorm and Apartment

Ideas generated in challenge 2 (Dorm and Apartment) were classified along the dimensions of “Type of Product”, “Primary Room or Location” and “Key Benefit Proposition” in the following categories:

Type of Product	***Primary Room or Location***	***Key Benefit Proposition***
Apparel/Accessories	Any	Convenience
Cleaning	Kitchen	Portability
Clocks, Watches, Alarms	Living	Multipurpose
Electronics/TV/Audio/computing	Bathroom	Hygiene
Food, Cooking, and Eating	Bedroom	Customization / Personalization
Furniture and Décor	Study / Office / Desk Area	Automation
Heating, Ventilation, Air Conditioning	Walls	Hi-tech
Lighting	Garden / Outdoors	Style
Personal Care and Health	Closet	Disposable
Power management and electricity		Reminder
Security		Safety
Storage		Value / Low Cost

For many practical problems, teams generate a number of possible solutions and then select a few for further investigation. We examine the effectiveness of two idea generation processes for such tasks— one, where the team works together as a team, and the other where individuals first work alone and then work as a team. We define effectiveness as the quality of the best ideas identified by the teams. We show that the quality of the best ideas depends on (1) the average quality of solutions generated, (2) the variance in the quality of generated solutions, (3) the number of solutions generated, and (4) the ability of the team to discern the quality of these solutions. We find that groups employing the hybrid process are able to generate more ideas, to generate better ideas, and to better discern their best ideas compared to teams that rely purely on group work. Moreover, we find that the frequently recommended brainstorming technique of building on each other's ideas is counter-productive: teams exhibiting such build-up neither create more ideas nor are the ideas that build on previous ideas better.

Summary of Changes for "Idea Generation and Quality of the Best Idea" by Girotra, Terwiesch and Ulrich

We would like to thank the AE and the two reviewers for their helpful comments on the previous version of this paper. We would also like to thank the DE for the opportunity to revise our work and, for granting us an extension on the resubmission deadline.

Based on the comments we received from the review team, we have engaged in a ***major*** revision of the paper. We have clarified our original theory, developed and tested new theory on the processes underlying our original observations, gathered new data, expanded our analysis, and improved the exposition of our work by relating it better to existing literature. Specifically, we have implemented the following major changes:

1. **New Ratings Methodology**: One of the major methodological concerns raised by the review team related to the idea evaluation process broadly, and specifically, to the metrics employed, and the fact that we used the same student population to evaluate the ideas that was previously involved in generating the ideas (AE-0, R1-7, R2-6, and R2-9).[1] To address these concerns, we collected three sets of new ratings data.
   a. We took the ideas generated as described in the paper (undergraduate design students) and used 41 MBA students enrolled in a course covering the financials of innovation to assess the business value of the idea.
   b. We conducted a market research study measuring the expressed purchase intent of 85 customers in the target market addressed by the ideas.
   c. We hired two research associates that scored each idea on multiple dimensions including (*technical feasibility, novelty, specificity*, *market demand*, and *overall value*. (R1-11a and R1-12abcd).

2. **Development of Theory**: The review team found that our theory took a quantitative approach and did not conform to standards from creativity literature. The theoretical foundation of our work is as much in stochastic models of product development and search as it is in creativity and brainstorming. For this reason, we submitted our work to the NPD department at Management Science, and not to a journal with a history of qualitative theory such as ASQ. We are also excited to see that our work co-evolves with some analytical work that is under review or recently has been published in Management Science. Nevertheless, we agree with the review team that for our paper the mathematical models reduce the potential readership of our paper. We now provide a unified block of theory that explicitly builds on the brainstorming literature as well as on the product development literature (AE-2). We do so by making all mathematical arguments qualitatively, referring to a set of modeling papers and where necessary, providing mathematical statements in the appendix. Figure 1 of the main paper illustrates our new theory. Again, we do want to point out that some of the disagreements with the reviewers might reflect that we just come from a different academic discipline, but we have done our best to work towards the ASQ standards, suggested by the review team (AE-6).

[1] Throughout this document, we use the notation Rx-y, to indicate comment number y from referee x.

3. **More micro level focus of our theory and analysis**: As was pointed out by the review team, our study has the potential to be the first that not just analyzes the outcomes of the idea generation process but also the process leading to these outcomes (AE-1b). One of the key challenges towards such a process level theory is to acknowledge that the generated ideas are not independent of each other (DE-1, AE-5, R1-2, R1-31). There exist a number of reasons for dependencies among ideas, including ideas building on each other (typical for good group work, R1-2), ideas overlapping with each other or being redundant (likely to happen if multiple individuals work on the same problem in isolation, R1-10b), and non-stationary idea quality levels (people getting tired or running out of good ideas). We follow the advice of the AE and now emphasize this process level in both, theory development and empirical analysis. To address idea overlap and redundancy, we eliminate all overlapping ideas. With respect to ideas building up on each other, we now develop a methodology to measure a proxy for the extent of buildup in the idea generation process. We then analyze the impact of this buildup on the performance of the creative problem solving exercise. Hypothesis 4, Sections 4.3 and 7 of the revised paper are entirely devoted to studying these effects. We also investigated non-stationarity in idea quality, but found no evidence for this non-stationarity.

4. **Refined and extended statistical analysis**: We have addressed the econometric issues raised by the review team (R1-6, R2-14, R2-16, R2-17) and derived a set of new results relating to the micro-level process of idea generation discussed above.

5. **Improved exposition**: We have completely rewritten the paper. This includes eliminating the mathematical discussion of our theory, strengthening our hypotheses development, an improved attempt at acknowledging the existing literature and providing a much more comprehensive explanation of our methodology. Specifically, we now provide a more detailed description of our experimental set-up, including the number of subjects, and the instructions that were given to raters and subjects. We have also clarified our concept of quality and made sure that both subjects and raters were in agreement on what constitutes high quality

In addition to these major changes, we have implemented a set of more detailed changes addressing each of your comments. They are summarized in the table following the bullet list. To be concise in our response, we use AE-x for the x-th comment from the AE and Ry-z for the z-th comment from Referee y.

Overall, we believe that we have a much stronger manuscript now than we had before. Thank you again for the constructive comments and for the opportunity to revise our work.

Raised by	Issue raised by the review team	Implemented Change/Comments
DE-1, AE-5, R1-2. R1-3	*Dependence in the quality of the ideas created* Independence is a starting point for almost any statistical model, it does not hold here. You are freer to look at this if you do not develop a formal model that requires the assumption of dependence. Dealing with dependence is hard, but even a "minor purchase" on this would be a "big deal" (AE-5). Extend your measures to not only best ideas but also whether ideas built on each other / abandon the idea of independence (R1-2); the observations of Sutton at IDEO suggest that there exists an order effect (hopefully with the last ideas being better); In other words, I would expect an order effect with later ideas having higher quality for the team design while the independence assumption might hold for the hybrid team. (R1-3)	Your comments identify a major weakness in majority of the prior experimental literature on brainstorming and our original manuscript: ideas generated in a brainstorming process are not like cars produced in an assembly line. Ideas are outputs of the brainstorming process at time *t* but then also become inputs for the process at time *t+1*. This creates dependences among ideas, including potential correlations in their quality levels. Our previous analysis indeed treated each observation as independent. As you point out, this is an incorrect statistical model of the idea generating process. Moreover, it also ignores a very exciting aspect of brainstorming – as observed in the Sutton and Hargadon study at IDEO, people working together, buildup and refine each other's ideas. There exist a number of reasons for dependencies among ideas, including ideas building on each other (typical for good group work, R1-2), ideas overlapping with each other or being redundant (likely to happen if multiple individuals work on the same problem in isolation, R1-10b), and non-stationary idea quality levels (people getting tired or running out of good ideas). We follow the advice of the AE (AE-1b) and now emphasize this process level in both, theory development and empirical analysis. This allows us to achieve much more than a "minor purchase" and we thank you for pushing us into this direction. Specifically, we first eliminate all redundant ideas from synthetic teams, as they might bias our results on the productivity of different idea generating processes. Next, we develop a methodology to measure a proxy for the extent of buildup in the idea generation process. We then analyze the impact of this buildup on the performance of the creative problem solving exercise.

		Hypothesis 4, Sections 4.3 and 7 of the revised paper are entirely devoted to studying these effects. We find evidence that teams do indeed build up more on each other's ideas but this buildup does not necessarily translate into substantial advantage over the hybrid process either in terms of having a larger pool of ideas to select, or in increasing the average quality of ideas. In fact, we find some evidence that ideas that build-up on each other tend to be systematically worse in terms of idea quality. We also investigated non-stationarity in idea quality, but found no evidence for this non-stationarity
AE-0	This will probably require another round of experiments	In response to the comments we received from the review team, we redid significant parts of our experiment and have expanded our data set. Specifically, we have collected new data along two dimensions: the idea evaluation or rating data, as well as classifying the content of ideas on a structured space. As far as the idea evaluation phase is concerned: a. We took the ideas generated as described in the paper (undergraduate design students) and used 41 MBA students enrolled in a course covering the financials of innovation to assess the business value of the idea. b. We conducted a market research study measuring the expressed purchase intent of 85 customers in the target market addressed by the ideas. c. We hired two research associates that scored each idea on multiple dimensions (including novelty and feasibility) (R1-11a, R1-12abcd). While we did not video tape the idea generation process, we had designed the experiment in a way that enabled us to analyze the idea generation process at the micro-level. This includes: a. Each idea had a sequence number attached to it that uniquely determines

		the ideas created before and after it. b. Since all the ideas in n instance of the idea generation process come from the same domain, it is possible to compare the similarity in content of any two ideas. c. This similarity allows us to measure the extent to which an idea builds up on a previously expressed idea. We use this micro-level data to first verify the assertion form existing literature, that teams do indeed buildup more on previously expressed ideas than the hybrid process. Next, we evaluate the impact of this buildup on different properties of the idea generating process. Specifically, we find that the more buildup in teams does not lead to advantages over the hybrid process, either in terms of the number of ideas generated or in increasing the average quality of ideas. Hypothesis 4, Sections 4.3 and 7 of the revised paper are entirely devoted to studying these effects.
AE-1a, R1-4, R2-4	Section 3 did not add much; derivations in Section 3 did not add much to the paper; why do you need the stylized facts / link to hypotheses is vague	The theoretical foundation of our work is as much in stochastic models of product development and search as it is in creativity and brainstorming. For this reason, we submitted our work to the NPD department at Management Science, and not to a journal with a history of qualitative theory such as ASQ. We are also excited to see that our work co-evolves with some analytical work that is under review or recently has been published in Management Science. Nevertheless, we agree with the review team that for our paper the mathematical models reduce the potential readership of our paper. We now provide a unified block of theory that explicitly builds on the brainstorming literature as well as on the product development literature (AE-2). We do so by making all mathematical arguments qualitatively, referring to a set of modeling papers and where necessary, providing mathematical statements in the appendix. Figure 1 of the main paper illustrates our new theory. Again, we do want to point out that some of the disagreements with the reviewers might reflect that we just come from a different academic discipline, but we have done our best to work towards the ASQ standards, suggested by the review team (AE-6).
AE-1b	You have the experimental set-up to observe	One of the key challenges towards such a process theory is to acknowledge

	these processes	that the generated ideas are not independent of each other (see point DE-1, AE-5, R1-2, R1-31 above). Once we had realized this independence violation, we started to explore the various forms of dependencies among the ideas (similarity, time stationary, overlap, see above), which forced us to articulate a theory of what is happening inside the black box of the brainstorming process. We then coded the process level data that you mentioned in AE-0 and derived a set of new results. (Hypothesis 4, Section 4.3 and 7 of the revised manuscript) Thank you for pushing us into that direction – we feel that this process level analysis is an additional, distinctive feature of our work relative to the prior experimental literature in this field.
AE-2	Outline a unified block of theory; keep that separate from the analysis and the results	We have completely rewritten the paper. This includes eliminating the mathematical discussion of our theory and strengthening our hypotheses development. We also keep this part of the paper separate from analysis and results as you requested. (See Section 3 and Figure 1)
AE-3	I agree with most of the issues the reviewers raise. Respond in a convincing setting.	The review team has provided us with a number of great suggestions and has raised an array of legitimate issues. In this document, we explain in great detail how we addressed every one of these 49 points.
AE-4	Assumptions about means – you focus too much on the variance while in practice the mean is really important as well; we have to control for mean effects	We agree with you that the mean is practically one (if not the most) important variable. For this reason, when we study differences in variance, we explicitly control for differences in mean in our econometric analysis by we introduce fixed (and random) effects at different levels- the idea creator level and the rater level. When we test for differences in the mean, we explicitly measure the mean effect arising out of treatment while controlling for the mean effect arising from other factors such as the creators abilities and/or the raters rating scheme. Our functional form allows for different parameters for mean and for variance and thus, we are able to identify both of these parameters. We have improved the presentation of our econometric analysis in the paper to make this more explicit.

AE-6	Your theory section is very thin, look at ASQ paper	Again (see AE-1a, R1-4, R2-4), we feel that stochastic models of problem solving fit well within the scope of this department at Management Science. Thus, it seems to us that some of the disagreements with the reviewers on theory development might reflect that we just come from a different academic discipline. Nevertheless, it lies in our interest to write this paper in a way that it has as large of a readership as possible – and this means that it has to be accessible and of interest to the brainstorming community. For this reason, we appreciate your help and have done everything we could to work towards ASQ standards as far as theory development is concerned.
AE-7, R1-1	Acknowledge the brainstorming literature more explicitly (AE-7) Frame the introduction more around the brainstorming literature (R1-1)	We have expanded the discussion of the brainstorming literature in the introduction. We also elaborate on the connection to the Innovation Management literature broadly and specifically to the new (and very active) area of problem solving in product development.
AE-8	Page 5, lines 23-29. This is hard to believe unequivocally	The statements in question do not exist in the paper any more.
AE-9	The pure collaborative treatment might be somewhat unrealistic, because in the real world, most people are doing hybrid.	The focus of our work is indeed the hybrid process and we use the pure collaborative process as a reference model for comparison. We agree with your observation that the pure collaborative process or the team process is somewhat rare in managerial settings (though we do believe that it does exist), yet it is (a) the best condition to test the ideas (b) the dominant approach followed in the brainstorming literature (see AE-7) and (c) the approach on which we have most theory available (AE-6). We have rewritten the paper and now explicitly acknowledge that managerial settings differ from the treatment provided in the lab.
AE-10	You overstate your results on page 18, page 20, and page 24	We now tone down this discussion and provide a cleaner explanation for our findings.
AE-11	Page 13, page 19, and page 21 are weak theory: pull back and convince the reader of a few interesting ideas rather than talking them through	Our findings related to the micro level process of idea generation make this part of the paper substantially more interesting. Rather than just reporting the results on outcomes, we can explain the process that generated the outcomes.

	results	In addition to discussing our process findings, we also relate our findings to the literature of search in product development (Sommer and Loch, Terwiesch and Loch). This literature distinguishes between different solution spaces (structured and unstructured) and the implications this has for a stream of ideas generated from this space.
AE-12	Selectively review some of the creativity literature	We have added some references to the creativity literature
R1-5	Hypothesis 3 in particular is not sufficiently motivated. Link to Christina Gibson's work on cognitive processes and Davis (1987) (this is the hypothesis on evaluation ability)	We have expanded our discussion on the self evaluation capability. We believe that this result is interesting and important and agree with you that it previously had not been sufficiently motivated. We have reviewed and used the work from Christina Gibson and Davis to build our theory on evaluation capabilities.
R1-6, R2-14	Clarify the sample size, the number of participating groups and the number of participants	We now provide a more detailed description of our experimental set-up, including the number of subjects, and the instructions that were given to raters and subjects. Specifically, we have: - 44 participating idea generators - 11 teams and 11 hybrid teams - 41 raters for the business value of the ideas, leading to 8950 observations (idea x rater) - 85 subjects that expressed their personal purchase intent for the product or service described by the idea, leading to 18841 observations (idea x subject)
R1-7	Where did the judges come from?	We took the ideas generated as described in the paper (undergraduate design students) and used 41 MBA students enrolled in a course covering the financials of innovation to assess the business value of the idea. We also conducted a market research study measuring the expressed purchase intent of 85 customers in the target market addressed by the ideas. Since product ideas targeted the college market, we recruited college students from across campus (mostly not associated with Wharton).
R1-8	Need to add descriptive statistics and correlation table	Table 2 in the revised manuscript provides the mean level of different measured variables. Our data set has mostly categorical variables, ratings,

		etc. It is not obvious to us, what kind of correlation table the referee is indicating. If the referee can clarify exactly what descriptive statistics are of interest, we would be happy to include them.
R1-9	Did you do any manipulation checks? (can you show that the two processes differed). Manipulation check would help to rule out some of the alternative explanations (R1-10a-c)	We personally observed the idea generation process and the difference between the hybrid process and the group process. The group process clearly operated as a group process – the entire time was spent on brainstorming product ideas with one person speaking at a time. The hybrid process started out with individual idea generation – no discussion / interaction existed during this time.
R1-10a	Alternative explanation: the group had to spend time to establish a group routine (unless you instructed them in brainstorming, which is not explained)	Both group and hybrid group might incur a fixed time to establish a group routine. If this time investment was significant, the hybrid group approach would be impacted more – after all, it has a shorter time period for the group to work together. However, we find the opposite: the hybrid was more productive. Note further that the students participating in the experiment were had almost completed a product design course. All students had been exposed to some design work and had received formal brainstorming training. We believe the subjects had a pretty clear idea about the routines in a brainstorming meeting.
R1-10b, R2-20	Alternative explanation: How did you account for overlap / similar ideas	Thank you for raising this point – this goes back to the independence assumption that was challenged by the AE and the DE (see above). As we now explore the micro level process of idea generation in even greater detail, we have operationalized the concept of similarity. Similarity is measured by evaluating to what extent idea n is similar to idea n-1 on one or several attributes (e.g. an MP3 holder for the treadmill is similar to an MP3 holder for weightlifting). For every idea, we can determine which idea was created by the same (hybrid) group immediately before (after). This allows us to analyze if and to what extent (and with what impacts on quality and productivity) group members build on each other's ideas.

		Further, when we create synthetic groups from the individual ideation part of the idea generation exercise, we eliminate completely overlapping/redundant ideas or ideas that refer to the same user need and same identified solution.
R1-10c	Alternative explanation: it is the more structured approach that leads to the higher productivity (Goldenberg et al 1999)	The hybrid approach leads to a significantly higher productivity. As we show, this is mainly driven by the substantial productivity gain during the individual phase, which eliminates the previously established weaknesses of group brainstorming such as production blocking. We agree with you that the added structure might be an additional benefit of the hybrid approach, but we find that the individual phase is the main driver. For this reason, we now discuss your point in the paper (including the reference that you provide), but we do not see this as a threat to our main contributions.
R1-11a	How do you define / measure quality? A lot of prior research suggests that quality is a multi-dimensional variable.	We now measure quality in two ways; both of them are significantly improved from the previous version of the paper. a. We took the ideas generated as described in the paper (by undergraduate design students) and used 41 MBA students enrolled in a course covering the financials of innovation to assess the business value of the idea. b. We conducted a market research study measuring the expressed purchase intent of 85 customers in the target market addressed by the ideas. To address the multi-dimensionality of quality, we also created a multi-dimensional quality scheme composed of five different metrics: Technical Feasibility (to what extent is the proposed product feasible to develop at a reasonable price with existing technology), Novelty (originality of the idea with respect to the unmet need and proposed solution), Specificity (the extent to which the idea included a proposed solution), Demand (reflecting market size and attractiveness), and Overall Value. To rate ideas on these dimensions, we recruited a team of two graduate students specializing in new product development and asked them to rate each idea on these dimensions on 10 point scale. We discarded all ratings where the two raters disagreed by

		more than 2 points. Looking at the remaining ratings, we found that the five dimensions were highly correlated. Factor analysis suggested using only one composite factor for the five metrics. Further, each of the metrics was highly correlated with business value and probability of purchase that we evaluated using larger panels. In light of this correlation, we will present our results using the business value and purchase probabilities.
R1-11b	Research by Reinig & Briggs (2006) suggests that the way you sum up multi-dimensions of quality matters	We did not sum up the multiple dimensions of quality – we asked the raters to provide a holistic evaluation of the idea. The (2nd year MBA) students were asked to assign financial values to the ideas and thus had to make judgments about an idea's demand as well as the cost it would take to produce it. To further address your concern about how to evaluate the multiple dimensions of quality, we have conducted a purchase intent study using customers from the target population of the products. Purchase intent studies are a widely accepted methodology in product development and in Marketing. Subjects in the study need to determine the expected utility they would obtain from purchasing the product and then translate this in their likelihood of purchase. They thus aggregate the multiple dimensions of utility in the mind of the consumers into a single outcome variable that matters for managers, the expected future sales. All of these methodological details were somewhat vague in the previous version of the paper– we now discuss them at length, Section 4.2 of the revised paper.
R1-12a R1-12b	What dimensions of quality did the judges use / what dimensions were the group told to use? How were the judges trained?	Students were instructed to generate ideas with a focus on the business value of idea to an existing retailer (IKEA in the case of dorm products, Eastern Mountain Sports in the case of sports products). The judges were instructed to evaluate the business value of the idea (same exact wording). In our purchase intent study, we asked the subjects (we do not want to call

		them judges): "How likely would you purchase this product if it were available at a retailer near you?" We completely left it to the subjects how to aggregate the various dimensions of their utility function. We followed the protocol of purchase intent testing as established in the standard product development text-books (e.g. Ulrich and Eppinger) Finally, we hired two doctoral students who were initially instructed to evaluate each idea on the dimensions: technical feasibility, novelty, market demand, and overall value of the idea. After discussing several hypothetical ideas with the students to determine how to assess each of these dimensions, we added a fifth dimension, idea specificity. This reflected the fact that some of the hypothetical ideas we had generated to train the two doctoral students were more specific than other. For example, compare the idea "MP3 holder made out of neoprene wrapped around the forehead" with the idea "really cool MP3 holder that can be used while running". The former idea is more explicit (specific) about ***how*** the product would address the need and hence is of potentially larger value to the company.
R1-12c	What were the teams told how they should rank the ideas?	The teams had the exact same instructions as the MBA raters: to generate ideas with a focus on the business value of idea to an existing retailer (IKEA in the case of dorm products, Eastern Mountain Sports in the case of sports products).
R1-12d	How did the judges compare to each other in the coding of quality; inter-rater agreements or rater idiosyncrasies	Our econometric analysis uses a rating as an observation. A rating reflects the raw quality of the idea, but also the subjective opinion of the rater. A regression with dummies (fixed effects) for the ideas shows that a large amount of the variance in rating can be explained by the quality of the ideas alone – thus, there exists a significant (agreed upon) idea effect. We also control for rater fixed effects (raters might differ in their average rating across all ideas and creator fixed effects (individuals may differ in their ideation ability). It lies in the nature of a purchase intent study that raters do not have to agree. Consider the example of a sports-bra, which is more likely to be purchased by a female subject compared to a male. The fact that our results carry over

		to our new purchase intent ratings suggests that few of our products were niche products that only appealed to a small sub-set of the population. For the inter-rater reliability analysis with large number of raters we follow the prescriptions from Gwet (2002), reporting Kappa and AC1 statistics for both business value and purchase intent (page 16). We find very strong inter-rater agreement between our different raters
R1-13	The test of the third hypothesis is meaningless unless we know how the groups were asked to rate their ideas (and if those instructions were in line with what the judges used)	As we explained above, the instructions were the same for those generating and those evaluating. Moreover, using our new purchase intent survey, we now obtain a rigorous estimate for the demand potential of an idea.
R2-1	I am not convinced that the results hold under real world conditions / the results reflect the experimental time constraints– since both hybrid and team have the same amount of time	The goal of our comparison between the hybrid and the team processes is to identify how organization can best use its manpower to generate creative solutions. Consequently, we feel that a fair comparison must consider the same number of man-hours in the two treatments. In other words, since we want to compare effectiveness of the two treatments, we want to use the same level of input, and we can then compare the level of output. With respect to the time limits being a binding constraint, in our observation of the experiment, we found that none of the generating units actually ran out of time. Typically the idea generation rate slowed down significantly towards to the end. Thus, the time limits imposed did not reflect a binding constraint in any fashion.
R2-2	Results are driven by the fact that the ratings that are the basis of comparison are obtained from individuals rather than from teams of raters, thus individual raters compare better.	In the context of innovation for new products, what matters the most is the potential market size of the product. This market size is influenced by individual purchase decisions made by market participants. To get a fair measure of the business value and market size of the ideas proposed, we use a purchase intent survey. We agree with the referee that the individual mature of this survey may be driving our results, but given that in the categories of products that we consider, real purchases are likely to be individual decisions, we feel an individual purchase intent survey is a fair metric to capture, what we really care about— the size of the market for the

		products.
R2-3	Run the experiment again and provide incentives (e.g. for self-rating accuracy); ease time constraints; provide more training in techniques	We thank the referee for these suggestions. We did indeed re-run this part of the experiment and now we use different measures for rating. (As explained above, purchase intent and business value). Purchase intent surveys are an established method for estimating market sizes in marketing literature, and we believe in the context of new product development they provide a very good metric for desirability of different products. We agree with the referee that implementing an incentive compatible scheme, such as a real market for product/ideas with budget constraints and real money would capture the incentives better. In addition to establishing the right market framework for capturing the value of money, we would further need to build some mechanism to capture the utility from acquisition of potential products that do not exist in any form. Establishing all these is hard, and we are in fact not aware of any study which has done this before. Nevertheless, we agree with the referee's concern and highlight this as a limitation of our results.
R2-5	An great version of the paper would take learnings thus far and design a new treatment, which would have it all- high mean quality, high variance, greater quantity and objectivity	We thank the referee for this suggestion. We agree that it would indeed be nice to create a treatment which would have all the benefits of team and hybrid. In this paper, we have studied the performance of two common treatments and provided a comparison and while this study provides some indications on the design of a new treatment, it remains a significant challenge to achieve all the desired properties in any one treatment. We defer tackling this challenge for future work.
R2-6	Concern about small sample size and a single experiment	Our sample size in this study is actually significantly higher than other studies. We achieve this by getting a very large number of raters from each of our ideas (we have more than a 100 raters and each of our ideas is examined by over 50 different individuals). Previous studies have typically employed a small number of raters (typically, 2). We agree with the referee that these results follow from one experiment, but we would like to clarify that within this experiment, there are two different

		ideation domains and further the within-subjects design of the experiment explicitly controls for individual effects. We believe these design features limit some of the concerns around basing our results from a single experiment.
R2-6a	Concern about the minimal absolute differences in the mean quality of the ideas (0.2 only.)	The quality advantage of the hybrid treatment is 0.25 units of Business Value and 0.35 units of purchase intent (significant at the 0.01% level for both business value and purchase intent). While this advantage might look small in absolute terms, such an absolute measurement can be misleading. Specifically, we measured idea quality and the differences in idea quality on a 10-point subjective rating scale. However, these do not necessarily map linearly onto the economic value of the ideas. Thus, effects which appear as marginal differences in our results may be of much higher or lower consequence in economic terms. This would be a function of the domain. For instance, while marginal differences in quality can make or break a new business venture, they may have little impact on innovation efforts aimed at internal process improvements (see Dahan, E. and H. Mendelson (2001) and Terwiesch, C. and K. T. Ulrich (2009) for more details on this nonlinear relationship). Further, we would like to emphasize that the mean absolute difference re not the only factor the drive our results, in fact it is difference in means, productivity, variance and evaluation ability that all come together to give the hybrid a significant advantage (3 times larger in absolute rating scales than the advantage from mean)
R2-7	Both hybrid and team method have pretty poor ability to rate the ideas, spearman correlations of 0.2	We agree with the referee's observation. Across treatments, the self evaluation ability is very small (and in some cases non-existent). We think this is one of our most salient findings- self evaluation abilities are generally pretty small. This has important implications on how organizations must design their idea generation and selection processes.
R2-8	Apparently individuals are better at rating their own ideas compared to their team members rating the individual's ideas	This is indeed correct, individuals rating their own ideas are better than a group of individuals rating the idea, where the group includes the original creator.
R2-9	The raters should be outside the group of test	We thank the referee for this important suggestion. We have implemented

	subjects	this change and we now use entirely distinct subject pools for idea generation and for idea evaluation.
R2-10	Page 9; lines 33-45 explain more that the upper tail and variance matter a lot.	We have entirely rewritten this section and we hope these points are better highlighted in the current version.
R2-11	Page 11, line 15-26 Where is the “fidelity” of the ranking process used later on in the paper?	We have entirely rewritten this section and we hope these points are better highlighted in the current version. Our new theory incorporates fidelity of the rating process directly.
R2-12	More rigorous development of H2; this is counter intuitive and demands a more rigorous explanation	We agree with the referee that Hypothesis 2 as stated in the original paper was indeed counter-intuitive and in fact on further reflection we felt that this could be argued wither way. Thus, we do not state this as a formal hypothesis any more.
R2-13	Why is the hybrid process more objective in terms of self-evaluations	From a statistical perspective we know that a process that has access to more independent, unbiased estimates of quality will be able to construct more accurate estimates of quality. There are two potential sources of bias and interdependence in the idea generation and selection process. First, if the same unit that created the idea is also asked to evaluate the idea, this unit may be biased in favor of its own ideas. Furthermore, ideas that for one reason or another garnered discussion time in the creation phase are made salient and therefore most likely to be perceived as high quality by the team members. These sources of bias are more prevalent in the team process than in the hybrid process. This is because in the hybrid process, the majority of ideas are likely to have been created during the individual phase and then evaluated by others in the group phase, reflecting independence between creators and evaluators. A second source of interdependence arises among group members in a team setting. Previous research has shown that team members affect one another’s perceptions, judgments and opinions (Gibson (2001), Stasser and Davis (1981), Zander and Medow (1963)). Detailed observation of the team cognitive processes has found that often “high-status” members dominate the discussion (Bandura (1997), Bartunek (1984), Davis, Bray and Holt (1977), Gibson (2001), Laughlin and Shippy (2006)). Because of these effects, we believe that the aggregation of information in teams will reflect

		interdependence among group members, and thus will not result in estimates of quality that are as good as those of the hybrid process. The evaluation process involves two factors, the amount of independent information brought to bear and the mechanism for aggregating that independent information. The team process suffers on both counts, less independent information is brought to bear and the aggregation mechanisms have the chance of being dominated by one or two individuals. Thus, the hybrid process is perhaps superior in evaluating ideas.
R2-15	Substitute the words "subject group" for teams on page 15, lines 4-8	Fixed.
R2-16	Page 15, top paragraph. This discussion would benefitted by a flow diagram showing how the 44 subjects went through testing, step by step	Thanks for this suggestion; we have now added a flow diagram for this. (Figure 2)
R2-17	Page 15, line 53. A brief discussion of how the Darwinator works would be appreciated	We have now added a flow diagram for the experiment and have added more explanation for our rating process. We have not added much more detail about our software platform, the Darwinator as in this version of the paper, we use multiple different rating methods, not all of which utilize the Darwinator.
R2-21	Page 18, line 13. "Whereas for the hybrid process"	We have rewritten the section.
R2-24	Page 20, line 34: use a "," instead of ";"	We have rewritten the section.
R2-25	Several references seem to be missing on EC8	Our original manuscript had two sets of references, some for the main paper and others for the electronic companion. We suspect that the referee only saw one of the two sets. Nevertheless, in the current version all references should appear.