Organizational Behavior and Human Decision Processes 154 (2019) 96–117

Contents lists available at ScienceDirect

Organizational Behavior and Human Decision Processes

journal homepage: www.elsevier.com/locate/obhdp

When silver is gold: Forecasting the potential creativity of initial ideas

Justin M. Berg

Stanford Graduate School of Business, Knight Management Center, 655 Knight Way, Stanford, CA 94305-7298, United States

ARTICLE INFO

Keywords:
Creativity
Idea evaluation
Idea selection
Creative forecasting
Creative cognition
Construal level

ABSTRACT

Past research on idea evaluation has focused on how individuals evaluate the creativity of finalized ideas. But idea evaluation is also important early in the creative process, when individuals must forecast the potential creativity of rough initial ideas as they decide which to develop. Using five experiments, this paper examines individuals' accuracy in forecasting the potential creativity of their initial ideas. Participants ranked the potential creativity of their initial ideas before developing them into final ideas. Results suggest that participants tended to under-rank their highest-potential idea. The initial idea that participants thought was their second best tended to actually be their best idea in the end. Broadly, the results suggest that creators exhibit myopia when forecasting the potential creativity of their initial ideas, leading them to overlook their most promising initial ideas. However, forecasting at a higher (more abstract) construal level helped participants identify their best initial idea.

1. Introduction

Bringing creative ideas to life is usually hard work, making creativity a risky investment (Sternberg & Lubart, 1991). In many popular accounts of successful innovations, creators trace their thinking back to a rough initial idea that they spent considerable time and energy developing into their final idea, often at the expense of pursuing other initial ideas. In 1937, Pablo Picasso was busy sketching plans for a painting titled *Artist's Studio* that was set to debut at the World's Fair in several weeks (Weisberg, 2004). Then he heard the tragic news that the Nazis had bombed Guernica, a city in his native Spain, and he was struck with the initial idea for a new painting based on the tragedy. Despite the work he had put into *Artist's Studio*, he scrapped it and focused on finishing *Guernica* for the World's Fair instead. This bet seemingly paid off, as *Guernica* became one of Picasso's most acclaimed paintings. In 1990, the first vision of Harry Potter popped into J.K. Rowling's mind on a delayed train ride. She spent the next five years mapping out the story and characters for all seven books while keeping various day jobs, presumably discarding other initial ideas for books she had as a longtime aspiring author (Rowling, 2016). In 1995, Larry Page was a doctoral student deciding on a dissertation topic. He reportedly considered ten different initial ideas, one of which was creating a search engine using the logic of academic citations. He ultimately decided to pursue this idea with Sergey Brin, and after about a year of further development, they had the first prototype of Google (Battelle, 2005).

In the many historical accounts of this sort, creators are cast as visionaries who recognized from the beginning which of their initial ideas deserved further investment. But how much foresight do individuals typically have about the potential creativity of their rough initial ideas? Ideas are creative when they are judged as both novel and useful (Amabile, 1996). Past research on idea evaluation has focused on how individuals evaluate the current creativity of finalized ideas: how novel and useful ideas are in their present form (e.g., Berg, 2016; Girotra, Terwiesch, & Ulrich, 2010; Mueller, Melwani, & Goncalo, 2012; Rietzschel, Nijstad, & Stroebe, 2010; Runco & Basadur, 1993). Less attention has been paid to forecasting potential creativity: how novel and useful initial ideas could become after they are fully developed.

In organizations, employees often tackle creative projects in which the goal is to produce a novel and useful idea, perhaps for a new product, service, system, or method (Harrison & Rouse, 2015; Obstfeld, 2012). Models of the creative process highlight how within a given project, individuals often generate and explore a number of initial ideas before they converge on one and develop it into a final idea (Amabile, 1996; Lubart, 2001). Past research suggests that this exploration stage of the creative process is critical, as initial ideas often have anchoring effects that shape how creative they may become as they are developed into final ideas (Berg, 2014; Diehl & Stroebe, 1987, 1991; Kornish & Ulrich, 2014; Ward, Smith, & Finke, 1999). Thus, to produce the most creative final ideas, generating a number of initial ideas is not enough; individuals must then be able to rank their highest-potential initial ideas above the rest. How accurate are individuals in ranking the potential creativity of their initial ideas?

E-mail address: jmberg@stanford.edu.

https://doi.org/10.1016/j.obhdp.2019.08.004
Received 24 September 2018; Received in revised form 19 July 2019; Accepted 18 August 2019

Scholars have debated whether individuals can have any foresight about the potential creativity of their initial ideas. In particular, the debate has centered on the "Blind Variation and Selective Retention" (BVSR) model of creativity. The BVSR model was first proposed by Campbell (1960) and has since been advanced and supported primarily by Simonton (1999a, 1999b, 2003, 2011, 2012). The BVSR model posits that creative ideas emerge from a Darwinian process of trial and error, characterized by "false starts and wild experiments" that ultimately produce an idea that is judged as creative in the domain (Simonton, 1999b, p. 197). Since the development of creative ideas is assumed to be inherently random and unpredictable, the BVSR model implies that it is impossible for individuals to predict the potential creativity of their initial ideas. From this view, the aforementioned anecdotes were about hindsight, not foresight—creators only realized which of their initial ideas were high-potential after they fully developed their ideas and gained external appreciation for them.

Critics of the BVSR model have advocated for a more systematic, predictable view of the creative process, allowing for individuals to have some degree of foresight about the potential creativity of their initial ideas (e.g., Gabora, 2007, 2011; Kozbelt, 2007, 2008; Mumford, 1999; Sternberg, 1998; Weisberg, 2004, 2015; Weisberg & Hass, 2007). These scholars suggest that although individuals cannot foresee the entirety of their final ideas at the start of the creative process, they may have a rough vision for what their initial ideas could become, enabling them to predict the potential creativity of their initial ideas with some degree of accuracy.

Both sides of this debate have marshaled evidence in support of their respective views. In support of the BVSR model, Simonton has amassed a large body of evidence demonstrating that creators rarely get better at predicting the success of their ideas throughout their careers, as their ratio of hits to misses stays consistent over time (Simonton, 1999a, 1999b, 2003, 2011). Critics of the BVSR counter with evidence that some creators' hit ratios do improve over their careers (Kozbelt, 2007, 2008), as well as evidence that interventions and external stimuli can predictably influence creativity (e.g., Mumford, 1999; Sternberg, 1998). Both sides have also found evidence for their opposing views in the same anecdotes and case studies, including the career of Thomas Edison (Simonton, 2015; Weisberg, 2015) and Picasso's aforementioned masterpiece *Guernica* (Gabora, 2011; Simonton, 2007; Weisberg, 2004).

Although this ongoing debate has helped clarify the merits and limitations of the BVSR model, the diverging arguments and evidence leave little consensus on individuals' capacity to accurately forecast the potential creativity of their initial ideas. This topic is not just of theoretical interest; it has important practical implications as well. If individuals are unable to identify their highest-potential initial idea, the optimal strategy may be to focus on quantity and develop a number of initial ideas to maturity. Otherwise, individuals risk rejecting their highest-potential idea in favor of an inferior initial idea, undermining the creativity of the final idea they ultimately develop. In contrast, if individuals are able to identify their highest-potential initial idea, then they can focus on quality and be more selective in deciding which of their initial ideas to further develop. Despite the important theoretical and practical implications, the question remains: are individuals able to identify their highest-potential initial idea?

In this paper, I draw on theories of creative cognition (Ward et al., 1999) and construal level (Trope & Liberman, 2010) to advance theory on individuals' capacity to accurately forecast which of their initial ideas holds the highest potential creativity. I test the proposed theory using five experiments in which participants ranked their initial ideas in terms of potential creativity before they developed them into final ideas. The results provide important theoretical insights on idea evaluation and selection, help integrate the two perspectives in the BVSR debate, and shed new light on the role of construal level in the creative process.

2. The myopia of forecasting potential creativity

Amabile's (1996) model of the creative process outlines four stages through which ideas are developed within creative projects: (1) problem or task identification (framing the creative project), (2) preparation (collecting relevant knowledge), (3) response generation (generating possible idea), and (4) response validation (evaluating possible idea). Within a given creative project, individuals may cycle through this process many times, each time generating a new or revised idea (stage 3) and evaluating it (stage 4), including vis-à-vis the other ideas they have already generated. At first, the ideas they generate and evaluate will be initial ideas. I define initial ideas as concepts that represent an incomplete but discernable response to a given creative project. Initial ideas are generated at the beginning of creative projects, each representing a different way to possibly address the creative project at hand. Initial ideas are understood by those generating them to be rough starting points on which to build more elaborate ideas. Initial ideas may remain as mental representations that are never communicated, or they may be captured and communicated externally with words or images. The key is that initial ideas are discernable concepts to the individual generating them, each representing a different possible path forward for addressing the creative project. Final ideas are the output at the end of the creative project, after an initial idea has been fully developed. Thus, final ideas are relatively detailed and complete concepts meant to be communicated and used in their current form.

The present research complements past work that has focused on earlier stages of creative projects. In particular, some past work has highlighted the importance of the content individuals gather and use in stage 2 of Amabile's model (preparation), before they generate an initial idea in stage 3 (e.g., Berg, 2014; Finke, 1996; Ward et al., 1999). Other work has demonstrated that individuals underestimate the number of creative ideas they will be able to generate (Lucas & Nordgren, 2015). Although this work did not distinguish between initial and final ideas, the focus was on relatively brief ideas with a similar level of development as the initial ideas that are the focus of the present research. This suggests that individuals may underestimate the number of different initial ideas they are able to generate, highlighting the importance of encouraging individuals to generate a number of different initial ideas early in creative projects (Diehl & Stroebe, 1987, 1991). Building on these past perspectives, the present research focuses on the phase after individuals have generated a set of different initial ideas. In particular, how accurate are individuals in ranking their highest-potential initial idea above the rest in their set? The hypotheses proposed below suggest that the answer to this question may depend largely on timing (see Fig. 1 for a visual summary).

2.1. The accuracy of individuals' predicted rankings in the short term

In the short term, individuals' rankings of the potential creativity of their initial ideas may be fairly accurate. However, following the logic of the availability heuristic (Tversky & Kahneman, 1973), their forecasts are likely to be anchored by the current form of their initial ideas, as it may be difficult or impossible to imagine what their ideas could become when fully developed, but the current state of their ideas will be relatively salient in their minds (Schwarz et al., 1991). Thus, individuals' rankings may mostly reflect how creative they think each initial idea is in its current state. Although this may be problematic in terms of predicting long-term potential, it may take some development time for ideas to substantially differ from their initial form. Until this happens, individuals' predicted rankings should be fairly accurate in reflecting the current creativity of their developing ideas.

In contrast to this argument, much of the prior research on idea evaluation and selection paints a bleak picture of individuals' capacity to accurately assess the creativity of their own ideas. In general, this past research suggests that individuals struggle to accurately evaluate

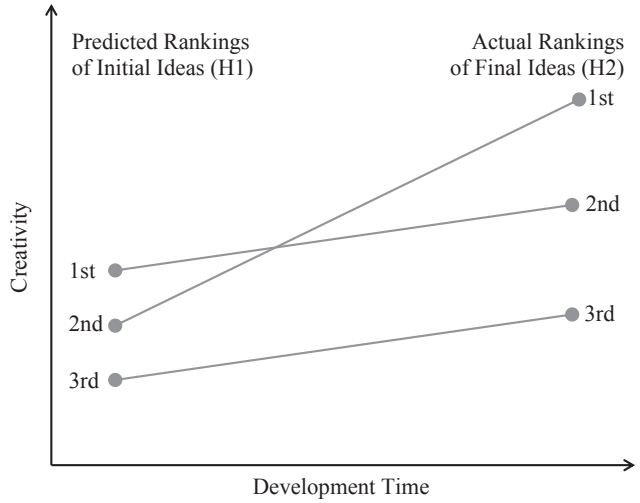

Fig. 1. Visual summary of Hypotheses 1 and 2.

creative ideas (e.g., Berg, 2016; Blair & Mumford, 2007; Licuanan, Dailey, & Mumford, 2007; Mueller et al., 2012; Rietzschel et al., 2010; Runco & Smith, 1992). However, the focus of the present research differs from this past work in three key ways that help explain why individuals' predicted rankings of their initial ideas may be fairly accurate in the short term.

First, prior work has largely focused on a different form of accuracy. Drawing on Moore and Healy (2008), at least two different forms of accuracy are relevant to idea evaluation and selection: estimation accuracy (evaluating the absolute creativity of any one idea), and placement accuracy (ranking how creative a set of ideas is compared with one another). Prior work speaks more to estimation than placement accuracy (e.g., Licuanan et al., 2007; Runco & Smith, 1992). This prior work has revealed important insights on what drives individuals to underestimate creative ideas, including feelings of uncertainty (Mueller et al., 2012) and thinking at a low construal level (Mueller, Wakslak, & Krishnan, 2014). Whereas estimation accuracy may be important in deciding whether to implement a given finalized idea or not, placement accuracy may be especially important early in creative projects, when individuals are trying to identify which of their initial ideas to pursue. Individuals could be quite far off in their absolute estimates for how creative their initial ideas will become, but still identify their most promising initial idea as long as their rankings are correct. Accurately ranking a set of initial ideas on creativity may be easier than accurately estimating the absolute creativity of each idea. Estimating absolute creativity requires individuals to figure out where a given idea ranks in the entire pool of ideas in the domain. This challenge is likely more difficult, complex, and susceptible to biases than ranking how a relatively small set of ideas compares with one another (Silvia, 2008).

Second, past research that does speak to placement accuracy has focused on relatively mature ideas, which are more complex and multifaceted than the relatively simple initial ideas that individuals may have early in creative projects (e.g., Blair & Mumford, 2007; Girotra et al., 2010; Rietzschel et al., 2010). Ranking relatively simple initial ideas on creativity may be easier than ranking more complex finalized ideas. Indeed, in a study on relatively simple ideas, Silvia (2008) found that individuals were fairly accurate in identifying their most creative ideas.

Third, past work has shown that when left to their own devices, individuals often have a bias against selecting novel ideas for implementation (Blair & Mumford, 2007; Mueller et al., 2012; Rietzschel et al., 2010). However, the focus of the present research is projects in which individuals have the explicit goal to be creative—i.e., produce a novel and useful idea. Past research suggests that the bias against novelty is mitigated when individuals are given the explicit goal to value novelty (Licuanan et al., 2007; Rietzschel et al., 2010; Shalley, 1991). Thus, the bias against novel ideas found in some prior research may not be relevant to assessments of initial ideas within projects that are explicitly focused on creativity.

In sum, when forecasting the potential creativity of their initial ideas, individuals' rankings may mostly reflect how creative they think each idea is in its present form (Tversky & Kahneman, 1973). On average, individuals should be able to accurately rank a relatively small set of initial ideas in terms of their current creativity (Silvia, 2008). Thus, individuals' predicted rankings will likely be accurate in the short term, at least until they have developed their ideas enough to substantially differ from their initial state.

Hypothesis 1. Individuals' predicted rankings of the potential creativity of their initial ideas are accurate in the short term.

2.2. The inaccuracy of individuals' predicted rankings in the long term

As individuals spend time developing their initial ideas, the accuracy of their predicted rankings may decrease. In particular, some of their initial ideas may hold more long-term potential than they predicted. What makes an initial idea high in long-term potential? Theories of creative cognition suggest that the most creative final ideas often begin as relatively abstract initial ideas (Finke, 1996; Ward et al., 1999). Because abstract initial ideas focus on general principles or concepts, rather than specific details, they allow individuals freedom to diverge from obvious or conventional ideas as they flesh out the details of their ideas (Förster, Friedman, & Liberman, 2004; Gabora, 2010; Kim & Zhong, 2017; Wiesenfeld, Reyt, Brockner, & Trope, 2017). In contrast, initial ideas that come to mind as relatively concrete are usually closer to obvious or conventional ideas, anchoring how creative the initial ideas are likely to become as they are elaborated (Berg, 2014; Gabora, 2010; Smith, 1995; Ward, 1994).

In practice, what this means is that as individuals generate initial ideas and consider which to pursue, their initial idea with the highest long-term potential is likely to be relatively abstract. However, in the short term, one of their relatively concrete initial ideas may be more creative. Concrete initial ideas usually come to mind more fully formed than abstract initial ideas (Gabora, 2010; Ward et al., 1999). Because concrete initial ideas emerge as more detailed and complete than abstract initial ideas, they likely begin closer to their ultimate potential than abstract initial ideas. Abstract initial ideas may require more development time to reach their potential in terms of novelty and usefulness, as more details need to be figured out to form a coherent and complete final idea. Before abstract initial ideas are fleshed out, it may be difficult for the creators of the ideas—and independent evaluators in the domain—to see much novelty or usefulness in them. In contrast, concrete initial ideas are likely to be more complete, detailed, and coherent, making it easier to see novelty and usefulness in them. Thus, while a relatively abstract initial idea may pave the way for the most creative final idea in the long term, in the short term, individuals' most creative initial idea is likely to be relatively concrete.

While concrete initial ideas are unlikely to change much from their initial form, it may be difficult or even impossible for individuals to predict all the ways in which abstract initial ideas may change as they are developed (Simonton, 2011). Because one of their relatively concrete initial ideas is likely to start out as more creative than their relatively abstract initial ideas, individuals may think that this concrete idea is their highest-potential idea (Schwarz et al., 1991). If they move forward with this relatively concrete initial idea, they may never realize that developing one of their more abstract initial ideas would have yielded a more creative final idea in the long term. In this way, the criteria individuals use to forecast the potential creativity of their initial ideas may make them myopic, as they may overvalue concrete initial ideas with relatively obvious short-term potential and undervalue abstract ideas with less obvious long-term potential. In turn, when

forecasting the potential creativity of their initial ideas, individuals are likely to rank a relatively concrete idea first, incorrectly favoring it over a more abstract idea that is actually their highest-potential idea in the long term.

Hypothesis 2. Development time moderates the accuracy of individuals' predicted rankings of the potential creativity of their initial ideas, such that individuals tend to under-rank their idea with the greatest long-term potential in favor of an idea that reaches its potential faster.

2.3. Additional theorizing on where individuals rank their most promising initial idea

The relatively concrete initial idea that individuals rank first may be the only idea they ultimately develop, which may leave behind a relatively abstract initial idea that was their most promising option because it had the highest long-term potential. Nonetheless, where the most promising initial idea falls in individuals' predicted rankings may have important theoretical and practical implications, as it speaks to their overall accuracy in forecasting potential creativity.

Individuals' most promising initial idea is likely to fall somewhere in the middle of their predicted rankings, reflecting a moderate degree of creativity compared to their other initial ideas. If an initial idea starts too low in creativity, it may not be able to improve enough to overtake initial ideas that started much higher in creativity. Initial ideas may improve in creativity as they are developed, and relatively abstract initial ideas may improve more than concrete initial ideas. However, prior research suggests that even relatively abstract initial ideas have a ceiling on how much they can improve in creativity. In studies that track ideas from their initial to final form, initial quality is significantly correlated with final quality (Berg, 2014; Kornish & Ulrich, 2014). Novelty may be more anchored by the initial idea than usefulness (Berg, 2014), but both dimensions may be anchored to some extent. It is unlikely that an initial idea that starts low in novelty and usefulness would be able to improve enough to overtake an initial idea that starts much higher, regardless of how abstract the initial idea may be.

Thus, individuals' most promising initial idea is likely to be moderately creative compared to their other initial ideas, meaning it should fall somewhere in the middle of their predicted rankings. For example, if individuals rank the potential creativity of three initial ideas, they may rank their most promising idea second, in between a relatively concrete idea ranked first that starts out as their most creative and their worst idea that they correctly rank third. These rankings should be accurate in the short term, as they reflect the creativity of the ideas in their initial form. However, in the long term, the concrete idea they ranked first is unlikely to improve much in creativity regardless of how much time is spent developing it. In contrast, the abstract idea they ranked second (which is actually their most promising) may improve more from development time and ultimately surpass the creativity of the concrete idea that they ranked first. Meanwhile, the initial idea they ranked third may start too far behind in creativity to surpass the other two, regardless of how abstract or concrete it may be.

To summarize, individuals' most promising initial idea will likely be moderately creative and relatively abstract, meaning it should fall somewhere in the middle of their predicted rankings and should be more abstract than the initial idea ranked first. When three initial ideas are generated and ranked, we would expect the most promising initial idea to be ranked second, reflecting that it is moderately creative compared to their other initial ideas. Where specifically the most promising initial idea falls in individuals' predicted rankings may depend on how many initial ideas they generate—the key is that it will likely fall in between first and last, reflecting a moderate degree of initial creativity. Thus, creative projects in which individuals generate three initial ideas provide the most parsimonious test of the proposed theorizing, which is why participants generated three initial ideas in Experiments 2–5 (Experiment 1 found similar results with four initial ideas). The first two hypotheses (H1 and H2) are tested in Experiments 1–4. Later, a third hypothesis is proposed on how individuals may overcome their myopic forecasts to identify their most promising initial idea, which is tested in Experiment 5.

3. Experiment 1: Testing H1 and H2 with open-ended development time

3.1. Participants and procedures

This first experiment included 208 participants in the U.S. recruited via Amazon MTurk (48.08% female, age 19–69, M_{age} = 33.91, SD_{age} = 9.78), who were each compensated $3.00. The objectives of Experiment 1 were to provide preliminary tests of H1 and H2, and to pilot how much development time to include in the manipulations for Experiment 2. To foster psychological realism (Berkowitz & Donnerstein, 1982), the task was in a domain with which participants were likely to have at least some familiarity: fitness equipment. Following evidence that setting the explicit goal to generate novel and useful ideas helps foster creativity (Rietzschel et al., 2010; Shalley, 1991), participants were informed that their goal was to develop a creative idea for a new piece of fitness equipment that is both novel and useful.[1]

The experiment unfolded in four main stages, mirroring the stages that are typically involved in the creative process for employees in organizations (Amabile, 1996; Lubart, 2001). First, participants were asked to generate four different initial ideas for a new piece of fitness equipment. This held the number of initial ideas constant across conditions, which was important given evidence on the value of considering a variety of initial ideas early in creative projects (Diehl & Stroebe, 1987, 1991), and that individuals may underestimate the number of initial ideas they are able to generate (Lucas & Nordgren, 2015). Second, participants were asked to rank their four initial ideas on potential creativity, which was explicitly defined as "how novel and useful your initial ideas could become after you spend time developing them into finalized ideas" (1st = most potential, 4th = least potential). Third, one of their four initial ideas was randomly selected for them to further develop and finalize. This created four conditions, one for each predicted rank (first, second, third, fourth). Participants were required to spend at least five minutes developing their randomly-selected initial idea. During this time, the survey had a blank field to serve as "scratch paper" so they could record notes on how they might develop their ideas. The survey would not advance until five minutes had passed, but participants were welcome to spend more time if they wanted. This enabled a pilot test of how much development time to include in the manipulations for Experiment 2. Participants spent an average of 10.25 min developing their ideas (SD = 4.50 min). Fourth, on the next survey page, participants described their final idea in a new blank field (the notes they wrote on the previous page were carried forward for their reference).

3.2. Creativity measure

To measure the creativity of participants' final ideas, I used the consensual assessment technique (Amabile, 1982, 1996). A separate sample of 450 consumers in the U.S. were recruited via MTurk to serve as independent raters (44.44% female, age 19–72, M_{age} = 33.38,

[1] To determine sample size, a medium effect size was used as a benchmark (f = 0.25). Power analysis showed that a sample size of 179 would provide 80% power in testing the interaction between development time and predicted rank (Faul, Erdfelder, Lang, & Buchner, 2007). To ensure this target was met, a total of 220 participants completed the survey, but 12 did not submit an actual final idea, and were thus omitted.

Table 1
ANCOVA model for Experiment 1.

	df	*F*	*p*	η^2
Rank	3	2.95	0.044	0.04
Time (in minutes)	1	3.58	0.060	0.02
Rank × Time	3	3.47	0.017	0.05
			$R^2 = 0.08$	
			$F(7, 200) = 2.54$, $p = .016$	

SD_{age} = 10.05), who were each compensated $3.00. Each rater assessed the creativity of 20 randomly-selected final ideas (out of 208 final ideas). Each final idea was rated by an average of 43.27 raters (*SD* = 2.01). To maintain internal consistency, raters were given the same broad definition of creativity as participants: "Overall degree to which the idea is both novel and useful," which they rated using a 7-point scale (1 = "extremely low", 7 = "extremely high"). Following suggested practice (Amabile, 1996), raters were asked to read ten ideas before they began rating to give them a frame of reference. An example of an idea rated high in creativity was a virtual reality hiking simulator (see Appendix A for examples of creative ideas from Experiments 1–5). Ideas rated low in creativity tended to be products that already existed with a minor feature added, such as a treadmill that plays music.

3.3. Results and discussion

Before testing H1 and H2 more precisely, the main effect of predicted rank was tested. One-way ANOVA showed the four predicted ranks marginally differed in final creativity, $F(3, 204) = 2.36$, $p = .072$, $\eta^2 = 0.03$. Post-hoc LSD tests showed that first-ranked ideas ($M = 4.03$, $SD = 0.76$) and second-ranked ideas ($M = 3.93$, $SD = 0.81$) did not significantly differ in final creativity, $p = .54$, $d = 0.13$, but first-ranked ideas were significantly higher in final creativity than third-ranked ideas ($M = 3.70$, $SD = 0.96$), $p = .044$, $d = 0.38$, and fourth-ranked ideas ($M = 3.66$, $SD = 0.88$), $p = .024$, $d = 0.45$. No other comparisons between the four ranks were significant.

To test H1 and H2, ANCOVA was used with predicted rank as a fixed factor and development time (in minutes) as a covariate. As expected, development time was a significant moderator of predicted rank (see Table 1 and Fig. 2), $F(3, 204) = 3.47$, $p = .017$, $\eta^2 = 0.05$. Consistent with H1, post-hoc LSD tests revealed that when participants spent one standard deviation below the mean in development time (5.75 min), first-ranked ideas finished significantly higher in creativity than second-ranked ideas, $t(200) = 2.44$, $p = .015$, $d = 0.36$, and third-ranked ideas, $t(200) = 2.22$, $p = .028$, $d = 0.31$, and were numerically—but not significantly—higher than fourth-ranked ideas, $t(200) = 1.37$, $p = .17$, $d = 0.19$. No other comparisons between ranks were significant. Consistent with H2, when participants spent one standard deviation above the mean in development time (14.75 min), second-ranked ideas finished significantly higher in creativity than first-ranked ideas, $t(200) = 2.05$, $p = .042$, $d = 0.29$, third-ranked ideas, $t(200) = 2.54$, $p = .012$, $d = 0.36$, and fourth-ranked ideas, $t(200) = 3.35$, $p = .001$, $d = 0.47$. No other comparisons between ranks were significant. In support of H1, these results suggest that participants' predicted rankings were mostly accurate in the short term. And in support of H2, the initial ideas that participants ranked second in potential creativity actually finished first in creativity, but only if they were developed for a relatively long time.

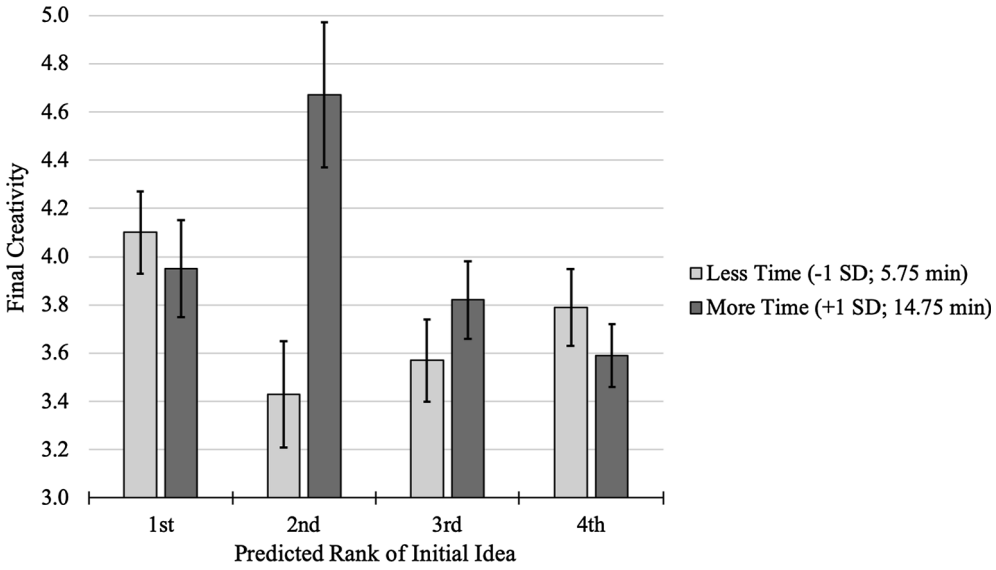

Fig. 2. Estimated marginal means from Experiment 1. Error bars are ± 1 SE.

4. Experiment 2: Testing H1 and H2 with development time manipulated

4.1. Participants and procedures

The second experiment included 300 participants in the U.S. recruited via MTurk (49.33% female, age 19–69, M_{age} = 34.22, SD_{age} = 10.39), who were each compensated $3.00. The experiment followed a 3 (predicted rank: first, second, third) X 2 (development time: less, more) between-subjects design.

The procedures built on Experiment 1 in three key ways. First, development time was manipulated; the durations were based on the distribution from Experiment 1 ($M = 10.25$ min, $SD = 4.50$). Specifically, less-time conditions were required to spend five minutes developing their randomly-selected initial idea, which was approximately one standard deviation below the mean for Experiment 1. More-time conditions were required to spend 15 min developing their randomly-selected initial idea, which was approximately one standard deviation above the mean for Experiment 1. Second, because the fourth initial idea was unnecessary for interpreting the results in Experiment 1, participants were asked to generate three initial ideas in Experiment 2. Third, to help ensure motivation was high across conditions,

Table 2
Means, SD's, interrater reliability, and correlations for Experiments 2–5.

	M	SD	ICC1	ICC2[a]	1	2	3	4	5	6
Experiment 2										
1. Initial Creativity	3.16	0.78	0.86	0.88-0.93						
2. Initial Novelty	3.45	0.96	0.89	0.90–0.94	0.92***					
3. Initial Usefulness	3.94	0.84	0.85	0.87–0.92	0.63***	0.44***				
4. Initial Abstractness	3.29	0.90	n/a	0.86	0.43***	0.38***	0.29***			
5. Final Creativity	3.78	0.77	0.90	0.91–0.95	0.54***	0.55***	0.25***	0.27***		
6. Final Novelty	3.84	0.80	0.85	0.91–0.94	0.54***	0.57***	0.20**	0.24***	0.96***	
7. Final Usefulness	3.87	0.65	0.85	0.88–92	0.37***	0.27***	0.45***	0.24***	0.64***	0.53***
Experiment 3										
1. Initial Creativity	3.12	0.79	0.88	0.85–0.94						
2. Initial Novelty	3.69	0.84	0.87	0.88–0.93	0.85***					
3. Initial Usefulness	3.23	0.91	0.90	0.85–0.95	0.93***	0.69***				
4. Initial Abstractness	3.91	0.90	n/a	0.72	−0.47***	−0.50***	−0.37***			
5. Final Creativity	3.81	0.74	0.86	0.87–0.96	0.56***	0.52***	0.53***	−0.24***		
6. Final Novelty	4.23	0.71	0.87	0.87–0.94	0.42***	0.50***	0.31***	−0.25***	0.85***	
7. Final Usefulness	3.60	0.84	0.91	0.85–0.96	0.57***	0.42***	0.61***	−0.23***	0.85***	0.54***
Experiment 4										
1. Initial Creativity	3.68	0.78	0.84	0.85–0.92						
2. Initial Novelty	3.78	0.85	0.87	0.86–0.94	0.93***					
3. Initial Usefulness	3.77	0.63	0.76	0.81–0.87	0.82***	0.68***				
4. Initial Abstractness	4.39	0.76	n/a	0.67	−0.19**	−0.14*	−0.23***			
5. Final Creativity	4.19	0.74	0.88	0.90–0.93	0.57***	0.58***	0.43***	0.03		
6. Final Novelty	4.31	0.79	0.89	0.90–0.94	0.57***	0.60***	0.40***	0.02	0.94***	
7. Final Usefulness	4.18	0.62	0.82	0.82–0.90	0.44***	0.39***	0.45***	−0.01	0.84***	0.72***
Experiment 5										
1. Initial Creativity	2.70	0.85	n/a	0.64						
2. Initial Novelty	2.82	0.92	n/a	0.64	0.86***					
3. Initial Usefulness	2.90	0.83	n/a	0.51	0.83***	0.78***				
4. Initial Abstractness	4.04	1.17	n/a	0.67	−0.48***	−0.43***	−0.45***			
5. Final Creativity	3.15	0.93	n/a	0.72	−0.04	−0.09	−0.08	0.15*		
6. Final Novelty	3.47	1.21	n/a	0.77	0.03	−0.01	−0.01	0.08	0.83***	
7. Final Usefulness	3.41	0.73	n/a	0.56	−0.02	−0.05	−0.02	0.07	0.74***	0.50***

Notes:
* $p < .05$.
** $p < .01$.
*** $p < .001$.
[a] When multiple groups of raters were used, the range of ICC2's is shown.

participants were told "If your final idea is considered very novel and very useful, you will be entered into a lottery for a $50 bonus." Following suggested practice, this reward was tied explicitly to the creativity of participants' ideas (Amabile, 1996).[2]

The number of words that participants generated in their "scratch paper" notes as they elaborated their ideas was used to test whether the more-time conditions actually led to more deliberate effort than the less-time conditions. Participants in the more-time conditions generated significantly more words ($M = 96.33$, $SD = 76.89$) than the less-time conditions ($M = 70.79$, $SD = 44.59$), $t(298) = 3.53$, $p < .001$, $d = 0.41$. This suggests that the time manipulations worked as intended in encouraging more (vs. less) development of initial ideas. The number of words did not significantly vary by predicted rank (or rank × time), hinting that participants were equally motivated regardless of the rank they were assigned.

[2] To determine sample size, the effect size of the interaction between development time and rank from Experiment 1 was used as a benchmark ($f = 0.23$). Power analysis showed that a sample size of 252 would provide 80% power in comparing the six conditions (Faul et al., 2007). Given the uncertainty in how manipulating development time might influence effect size, the target of 300 participants was set, which also enabled equally-sized groups of ideas for raters. In the first 300 completed surveys, 14 participants did not submit an actual final idea, and were thus replaced to reach the target of 300 participants (each condition had 1–3 of the 14 errant responses).

4.2. Measures

4.2.1. Creativity

The consensual assessment technique was again used to measure creativity (Amabile, 1996). Participants' initial and final ideas were assessed on overall creativity, using the same definition and scale as Experiment 1. A separate sample of 300 consumers in the U.S. were recruited via MTurk to serve as raters (51.33% female, age 19–73, $M_{age} = 34.92$, $SD_{age} = 10.28$). They were each compensated $4.50. To complement the consumer ratings, expert ratings of the final ideas were also collected. Fifty fitness experts were recruited via Qualtrics, including personal trainers, health club managers, fitness course instructors, athletics coaches, and physical education teachers (68.00% female, age 25–72, $M_{age} = 44.98$, $SD_{age} = 13.43$). They averaged 12.76 years of professional fitness experience ($SD = 10.14$, range: 2–40 years).

To enable interrater reliability statistics, the ideas were divided into equally-sized groups. Participants' 900 initial ideas were randomly divided into five groups of 180 ideas, and the 300 final ideas were randomly divided into five groups of 60 ideas. Each group of initial ideas was rated by 30 consumers, and each group of final ideas was rated by 30 consumers and ten experts (40 raters total). The order of ideas was randomized for each rater, and raters reviewed a random subset of ideas before they started rating (10 for final ideas, 15 for initial ideas). Past research suggests that novelty and usefulness may be weighted differently in assessments of overall creativity depending on the task/domain (Berg, 2014; Rietzschel et al., 2010). To shed light on the weights placed on novelty and usefulness in overall creativity, raters

Table 3
Means and SD's by condition for Experiment 2.

Condition	n	Final Creativity	Final Novelty	Final Usefulness
Less Time (5 mins.):				
1st Rank	51	3.75 (0.82)	3.82 (0.89)	3.86 (0.71)
2nd Rank	52	3.66 (0.70)	3.68 (0.72)	3.93 (0.64)
3rd Rank	49	3.56 (0.71)	3.70 (0.78)	3.67 (0.69)
More Time (15 mins.):				
1st Rank	50	3.83 (0.66)	3.85 (0.64)	4.01 (0.64)
2nd Rank	50	4.23 (0.89)	4.18 (0.91)	3.96 (0.59)
3rd Rank	48	3.65 (0.68)	3.80 (0.73)	3.77 (0.58)

also rated each dimension separately, using the same 7-point scale. Novelty was defined as "Degree to which the idea is unique from existing ideas." Usefulness was defined as "Degree of value offered by the idea." The consumer and expert ratings were highly correlated ($r = 0.68$ for creativity, $p < .001$), and results were similar when they were combined or analyzed separately, and thus the combined results were used (Appendix B reports the minor differences between consumers and experts). Each set of raters met conventional standards for interrater reliability (LeBreton & Senter, 2008). See Table 2 for interrater statistics across Experiments 2–5.

4.2.2. Initial abstractness

To help test the proposed theorizing, the abstractness (vs. concreteness) of participants' initial ideas was also measured. A separate set of ten independent raters were recruited from MTurk (40.00% female, age 23–36, $M_{age} = 29.50$, $SD_{age} = 4.90$), who were each compensated $50.00. Because assessing abstractness required relatively nuanced judgements, the same set of ten raters assessed all 900 initial ideas. Abstractness was defined as "Degree to which the idea involves general concepts or principles, as opposed to concrete or tangible details," which raters assessed on the same 7-point scale used for the creativity measures (1 = "extremely low", 7 = "extremely high"). Raters reviewed 15 ideas before they began rating. An example of an initial idea rated high in abstractness was "…a gravity or wind machine. It would act like how you go into a pool with a current. You would be able to direct the gravity or wind to provide resistance to a certain direction…" (see Appendix A for how this initial idea was developed). An example of an initial idea rated low in abstractness was "shoes with springs."

4.3. Results and discussion

4.3.1. Initial creativity

Consistent with H1, results showed that participants' predicted rankings were accurate in terms of initial creativity, meaning how creative their initial ideas were before any further development. Repeated-measures ANOVA showed that the three ranks differed in initial creativity, $F(2, 598) = 26.36$, $p < .001$, $\eta^2 = 0.08$, novelty, $F(2, 598) = 20.15$, $p < .001$, $\eta^2 = 0.06$, and usefulness, $F(2, 598) = 6.93$, $p = .001$, $\eta^2 = 0.02$. Post-hoc LSD tests showed that first-ranked ideas ($M = 3.29$, $SD = 0.79$) were significantly more creative than second-ranked ideas ($M = 3.17$, $SD = 0.76$), $p = .002$, $d = 0.18$, and second-ranked ideas were significantly more creative than third-ranked ideas ($M = 3.01$, $SD = 0.77$), $p < .001$, $d = 0.24$. In terms of initial novelty, first-ranked ideas ($M = 3.61$, $SD = 0.93$) were significantly more novel than second-ranked ideas ($M = 3.47$, $SD = 0.94$), $p = .006$, $d = 0.17$, and second-ranked ideas were significantly more novel than third-ranked ideas ($M = 3.28$, $SD = 0.98$), $p < .001$, $d = 0.21$. In terms of initial usefulness, first-ranked ideas ($M = 4.02$, $SD = 0.82$) were numerically—but not significantly—more useful than second-ranked ideas ($M = 3.96$, $SD = 0.83$), $p = .19$, $d = 0.07$, and second-ranked ideas were significantly more useful than third-ranked ideas ($M = 3.85$, $SD = 0.84$), $p = .017$, $d = 0.13$. These results are consistent with H1, in that participants' predicted rankings accurately reflected the creativity of their ideas in their initial form.

4.3.2. Final creativity

One-way ANOVA showed that the six conditions significantly differed in final creativity, $F(5, 294) = 5.12$, $p < .001$, $\eta^2 = 0.08$. See means by condition in Table 3 and Fig. 3. Two-way ANOVA showed a significant interaction between predicted rank (first, second, third) and development time (less, more), $F(2, 294) = 3.56$, $p = .029$, $\eta^2 = 0.02$. The main effects were also significant for rank, $F(2, 294) = 5.14$, $p = .006$, $\eta^2 = 0.03$, and development time, $F(1, 294) = 8.27$, $p = .004$, $\eta^2 = 0.03$. Consistent with H2, post-hoc LSD tests revealed that the "second-more" condition, meaning the condition in which participants developed the initial idea they ranked second for relatively more time, was significantly higher in final creativity than the other five conditions: first-less, $p = .001$, $d = 0.56$, first-more, $p = .008$, $d = 0.51$, second-less, $p < .001$, $d = 0.71$, third-less, $p < .001$, $d = 0.71$, and third-more, $p = .001$, $d = 0.83$. No other comparisons

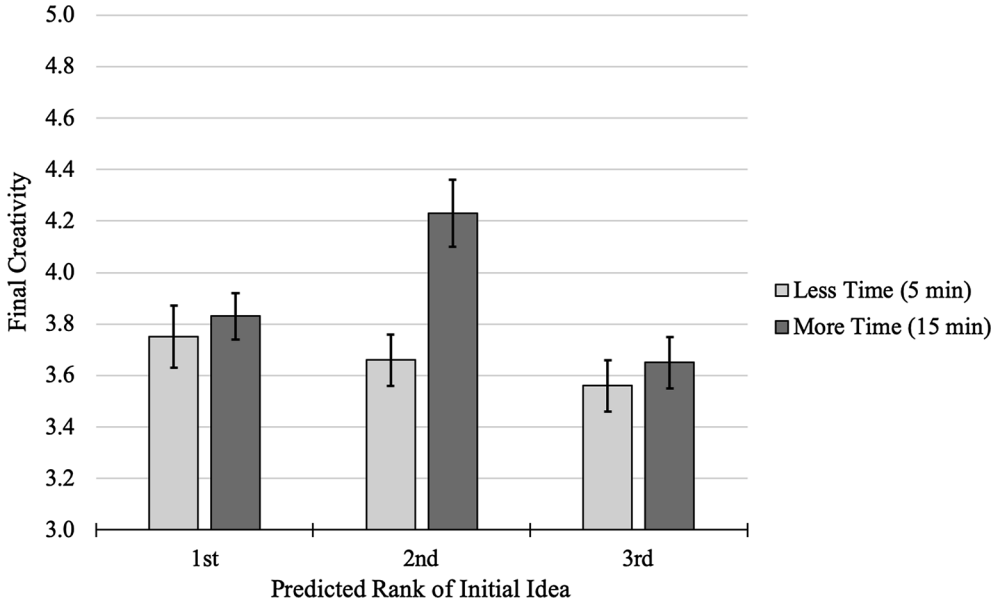

Fig. 3. Means from Experiment 2. Error bars are ± 1 SEM.

between the six conditions were significant.

These results for final creativity appear to be driven more by novelty than usefulness. The six conditions significantly differed in final novelty, $F(5, 294) = 2.73$, $p = .020$, $\eta^2 = 0.04$, but the differences were marginal for final usefulness, $F(5, 294) = 1.88$, $p = .097$, $\eta^2 = 0.03$. Post-hoc LSD tests for final novelty were similar to the pattern for final creativity. Specifically, the second-more condition was significantly higher in final novelty than the other five conditions: first-less, $p = .021$, $d = 0.40$, first-more, $p = .033$, $d = 0.42$, second-less, $p = .001$, $d = 0.62$, third-less, $p = .002$, $d = 0.57$, and third-more, $p = .016$, $d = 0.46$ (see means in Table 3). No other comparisons between the six conditions were significant.

Similar to Experiment 1, these results suggest that the initial idea participants thought was their second best in terms of potential creativity was actually their best, but only if they spent a relatively long period of time developing it. Also, consistent with H1, in the three less-time conditions, the final ideas scored in the order that participants predicted in their initial idea rankings, although these differences were not significant. The results also suggest that development time only mattered for the initial idea that participants ranked second. Initial ideas ranked first or third were not significantly affected by development time, while spending more time on initial ideas ranked second produced a significant boost in final creativity. This may be explained by the notion that second-ranked initial ideas were more abstract than first- and third-ranked initial ideas, which is discussed in the section that follows.

4.3.3. Initial abstractness

Consistent with the proposed theorizing, the initial ideas that participants ranked second tended to be more abstract than those ranked first. Repeated-measures ANOVA showed that the three ranks significantly differed in initial abstractness, $F(2, 598) = 5.84$, $p = .003$, $\eta^2 = 0.02$. Post-hoc LSD tests showed that second-ranked initial ideas ($M = 3.41$, $SD = 0.91$) were significantly more abstract than initial ideas ranked first ($M = 3.18$, $SD = 0.89$), $p = .001$, $d = 0.19$, and third ($M = 3.23$, $SD = 0.89$), $p = .013$, $d = 0.14$. First- and third-ranked ideas did not significantly differ, $p = .52$, $d = 0.04$.

Furthermore, ANCOVA was used to test whether abstract initial ideas improved more from development time than concrete initial ideas (see Table 4). In a model controlling for initial creativity, there was a significant interaction between initial abstractness and development time (more vs. less) in predicting final creativity, $F(1, 295) = 4.38$, $p = .037$, $\eta^2 = 0.02$. The pattern of results for this interaction suggests that abstract initial ideas improved more from development time than concrete initial ideas (see Fig. 4). When initial ideas were relatively abstract (one standard deviation above the mean in abstractness), more development time ($M = 4.00$, $SE = 0.08$) led to significantly greater increases in final creativity than less development time ($M = 3.63$,

Table 4
ANCOVA models for Experiment 2 testing whether abstract initial ideas improved more from development time than concrete initial ideas.

	Without Controlling for Initial Creativity				Controlling for Initial Creativity			
	df	*F*	*p*	η^2	*df*	*F*	*p*	η^2
Initial Abstractness	1	22.68	< 0.001	0.07	1	0.48	0.49	0.002
Time (less vs. more)	1	9.57	0.002	0.03	1	8.21	0.004	0.03
Initial Abstractness × Time	1	2.20	0.14	0.01	1	4.38	0.037	0.02
Initial Creativity (control)					1	95.00	< 0.001	0.24
	$R^2 = 0.11$				$R^2 = 0.33$			
	$F(3, 296) = 11.83$, $p < .001$				$F(4, 295) = 35.44$, $p < .001$			

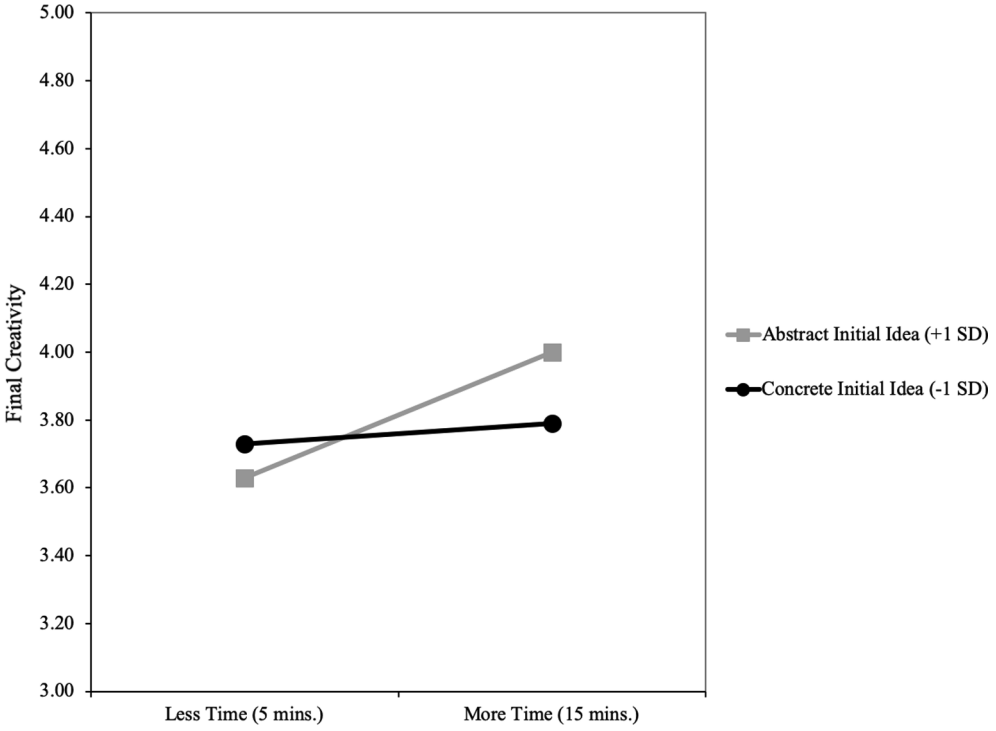

Fig. 4. Estimated marginal means from Experiment 2 for interaction between initial abstractness and development time.

$SE = 0.08$), $t(295) = 3.51$, $p = .001$, $d = 0.41$. In contrast, when initial ideas were relatively concrete (one standard deviation below the mean in abstractness), more development time ($M = 3.79$, $SE = 0.07$) did not significantly change final creativity compared to less development time ($M = 3.73$, $SE = 0.08$), $t(295) = 0.54$, $p = .59$, $d = 0.06$.

These results are consistent with the proposed theorizing, in that participants tended to rank a relatively concrete initial idea first, undervaluing the relatively abstract idea they ranked second. In general, the results suggest that individuals have some degree of foresight about the potential creativity of their initial ideas, but the myopic nature of this foresight may lead them to overlook their most promising initial idea.

5. Experiment 3: Testing H1 and H2 with a mixed factorial design

5.1. Participants and procedures

Participants included 294 students and staff at a large university on the U.S. West Coast (68.37% female, age 18–66, $M_{age} = 25.90$, $SD_{age} = 8.68$), who were each compensated $22.00.[3] Experiment 3 utilized a mixed factorial (within- and between-subjects) design, which built on the between-subjects designs in Experiments 1 and 2 in two key ways.

The first way is that participants developed and finalized the top two initial ideas (out of three) in their predicted rankings of potential creativity. They worked on their two ideas separately; they finalized and submitted one idea and then finalized and submitted the other one. Whether participants started with their first- or second-ranked idea was randomized. This helped rule out potential confounds of randomly assigning participants only one of their initial ideas. For example, it is possible that participants in Experiments 1 and 2 who were assigned their second-ranked ideas were compelled to work harder in developing their ideas than those assigned their first-ranked ideas.

To avoid biasing participants' predicted rankings, they were not told which idea(s) they would be developing until after submitting their predicted rankings. Participants were then informed that the goal was to make both of their top two ideas as creative as possible. To reinforce this goal, participants were informed: "The creativity (novelty and usefulness) of your two final ideas will be rated by a large group of consumers. The ratings for your two ideas will then be averaged together. If your average is in the top 10% of participants, you will receive a $10 bonus." Development time was manipulated between subjects, such that participants were either encouraged to spend five minutes (less time) or 15 min (more time) developing each of their top two initial ideas. Participants in the less-time condition completed a filler task after submitting their two final ideas so that all participants were compensated for an approximately equal amount of time.

The second important way that Experiment 3 complemented Experiments 1 and 2 is that a different task was used, to help rule out the possibility that idiosyncrasies with the fitness equipment task drove the results in Experiments 1 and 2. In particular, one potential limitation of the fitness equipment task is that a lot of fitness equipment already exists. As a result, participants were likely to have relatively rich, concrete schemas come to mind, making their initial ideas relatively concrete on average (Schwarz et al., 1991; Tversky & Hemenway, 1984). Results may vary when participants' initial ideas are more abstract on average. Thus, the task in Experiment 3 was more novel so that participants had less concrete schemas associated with it. In particular, the task was generating creative ideas for keeping people alert and engaged while using self-driving cars. All participants read the same introduction to the task, which explained that until self-driving cars become reliably autonomous, it will be important for passengers to stay engaged so they can take over driving if need be.

To further facilitate the abstractness of participants' initial ideas, I drew on procedures used in Berg (2014). When generating their three initial ideas, participants were told "It is helpful to use aspects of existing ideas as inspiration for coming up with new ideas. So, when coming up with your different ideas, start each idea by combining aspects of the two randomly-selected concepts below." All participants were shown photos of the same two objects/concepts (but in random order): one object associated with automobiles (a cup holder) and one from another domain (a slot machine). The logic of this approach is that combining elements of these previously disparate concepts required participants to think analogically—and thus more abstractly—facilitating the abstractness of their initial ideas (Dahl & Moreau, 2002; Gentner, 1989; Thompson, Gentner, & Loewenstein, 2000).

Like Experiment 2, the number of words that participants generated in their "scratch paper" was used as a manipulation check. Participants in the more-time condition generated significantly more words ($M = 95.73$, $SD = 84.31$) than the less-time condition ($M = 67.02$, $SD = 52.39$), $t(586) = 4.98$, $p < .001$, $d = 0.41$. This suggests that the time manipulations worked as intended in encouraging more (vs. less) development of initial ideas. The number of words did not significantly vary by rank (or rank X time), hinting that motivation was not affected by rank.

5.2. Measures

5.2.1. Creativity

The overall creativity, novelty, and usefulness of participants' initial and final ideas were measured using the same procedures as Experiment 2. A separate sample of 630 consumers in the U.S. were recruited via MTurk to serve as raters (54.44% female, age 18–82, $M_{age} = 38.26$, $SD_{age} = 12.06$), who were each compensated $5.00. Participants' 870 initial ideas were randomly divided into nine groups of 98 ideas, and their 588 final ideas were randomly divided into 12 groups of 49 ideas. Each group was rated by 30 consumers.

To complement the consumer ratings, six graduate students in Mechanical Engineering served as expert raters for the final ideas. They had all completed a course on self-driving car technology approximately two weeks before they started their ratings. The six experts each rated all 588 final ideas. Thus, each initial idea was rated by 30 consumers, and each final idea was rated by 30 consumers and six experts (36 raters total). The consumer and expert ratings were highly correlated ($r = 0.77$ for creativity, $p < .001$), and results were similar when they were combined or analyzed separately, and thus the combined results were used (Appendix B reports the minor differences between consumers and experts). Each set of raters met conventional standards for interrater reliability—see Table 2 (LeBreton & Senter, 2008). An example of a final idea rated high in creativity was an interactive game in which passengers are rewarded for identifying potential dangers on the road (see Appendix A for full idea—the corresponding initial idea was also rated high in abstractness). An example of a final idea rated low in creativity was having the car make coffee to keep passengers awake (the corresponding initial idea, "have car make coffee," was rated low in abstractness).

5.2.2. Initial abstractness

The abstractness of participants' 870 initial ideas was measured using the same procedures as Experiment 2 by an independent set of ten raters recruited from MTurk (20.00% female, age 24–51, $M_{age} = 33.90$, $SD_{age} = 7.23$), who were each paid $50.00.

[3] The original target was 300 participants, to maintain consistency with Experiment 2. But after removing participants who did not submit actual final ideas, the count (N = 294) allowed for equally divisible rater groups (see *Measures* section), and thus no additional responses were collected.

5.3. Results and discussion

5.3.1. Initial creativity

Consistent with H1, results showed that participants' predicted rankings were accurate in terms of initial creativity. Repeated-measures ANOVA showed that the three ranks differed in initial creativity, $F(2, 586) = 42.85$, $p < .001$, $\eta^2 = 0.13$, novelty, $F(2, 586) = 114.74$, $p < .001$, $\eta^2 = 0.28$, and usefulness, $F(2, 586) = 74.86$, $p < .001$, $\eta^2 = 0.20$. Post-hoc LSD tests showed that first-ranked ideas ($M = 3.31$, $SD = 0.75$) were significantly more creative than second-ranked ideas ($M = 3.13$, $SD = 0.77$), $p < .001$, $d = 0.25$, and second-ranked ideas were significantly more creative than third-ranked ideas ($M = 2.90$, $SD = 0.81$), $p < .001$, $d = 0.29$. In terms of initial novelty, first-ranked ($M = 4.23$, $SD = 0.66$) and second-ranked ideas ($M = 4.20$, $SD = 0.79$) did not significantly differ, $p = .59$, $d = 0.04$, but first-ranked ideas were significantly more novel than third-ranked ideas ($M = 3.53$, $SD = 0.86$), $p < .001$, $d = 0.75$, and second-ranked ideas were significantly more novel than third-ranked ideas, $p < .001$, $d = 0.70$. In terms of initial usefulness, first-ranked ($M = 3.61$, $SD = 0.81$) and second-ranked ideas ($M = 3.58$, $SD = 0.89$) did not significantly differ, $p = .61$, $d = 0.03$, but first-ranked ideas were significantly more useful than third-ranked ideas ($M = 3.01$, $SD = 0.93$), $p < .001$, $d = 0.63$, and second-ranked ideas were significantly more useful than third-ranked ideas, $p < .001$, $d = 0.58$. These results provide additional support for H1 in terms of overall creativity, and partially support H1 in terms of novelty and usefulness separately. The fact that first- and second-ranked initial ideas significantly differed in overall creativity, but not in novelty and usefulness separately, suggests that to be deemed creative in Experiment 3, ideas needed a balance of both novelty and usefulness (this was the case with final ideas as well, which is discussed below).

5.3.2. Final creativity

In Experiment 3, participants finalized their top two initial ideas. Because they submitted one final idea before working on their next idea, the most direct replication of Experiments 1 and 2 was examining the final ideas that participants submitted first. The order in which participants developed their first- and second-ranked initial ideas was randomized, and development time was manipulated (less vs. more), meaning there were four conditions regarding the first set of final ideas submitted. These four conditions significantly differed in final creativity, $F(3, 290) = 4.87$, $p = .003$, $\eta^2 = 0.05$. Consistent with Experiments 1 and 2, post-hoc LSD tests showed that the second-more condition ($M = 4.19$, $SD = 0.80$) was significantly higher in final creativity than the other three conditions: first-less ($M = 3.87$, $SD = 0.65$), $p = .005$, $d = 0.44$, first-more ($M = 3.92$, $SD = 0.56$), $p = .018$, $d = 0.39$, and second-less ($M = 3.78$, $SD = 0.71$), $p < .001$, $d = 0.54$. No other comparisons between conditions were significant.

To account for the fact that each participant submitted two final ideas, a mixed-effects model was used (see Table 5). The model included a by-participant random intercept, and rank (first, second) and development time (less, more) were fixed factors. The interaction between rank and development time was significant, $F(1, 584) = 11.68$, $p = .001$. The main effect was not significant for rank, $F(1, 584) = 0.002$, $p = .96$, and was marginal for development time, $F(1, 584) = 3.44$, $p = .064$. Post-hoc results regarding the significant interaction were consistent with H1 and H2 (see Fig. 5). Specifically, paired-samples t-tests showed that when participants spent less time developing their top two initial ideas, first-ranked ideas ($M = 3.82$, $SD = 0.69$) finished significantly higher in creativity than second-ranked ideas ($M = 3.67$, $SD = 0.78$), $t(149) = 2.48$, $p = .010$, $d = 0.21$. Conversely, when participants spent more time developing their top two initial ideas, second-ranked ideas ($M = 3.96$, $SD = 0.84$) finished significantly higher in creativity than first-ranked ideas ($M = 3.81$, $SD = 0.63$), $t(143) = 2.25$, $p = .026$, $d = 0.20$. These results provide additional support for H1 and H2.[4]

Table 5
Mixed-effects model for Experiment 3.

	df	*F*	*p*
Rank (2 vs. 1)	1	0.002	0.96
Time (less vs. more)	1	3.44	0.064
Rank × Time	1	11.68	0.001
	$F(3, 584) = 5.05$, $p = .002$		

The differences in overall creativity appear to be driven by a combination of novelty and usefulness together, as the results for both novelty and usefulness were numerically in the predicted direction but not significant. Specifically, when participants spent less time, their first-ranked ideas finished numerically—but not significantly—higher in novelty ($M = 4.22$, $SD = 0.67$) and usefulness ($M = 3.60$, $SD = 0.81$) as compared to second-ranked ideas in novelty ($M = 4.13$, $SD = 0.77$), $t(149) = 1.40$, $p = .16$, $d = 0.12$, and usefulness, ($M = 3.48$, $SD = 0.86$), $t(149) = 1.88$, $p = .063$, $d = 0.15$ (the difference was marginal for usefulness). Conversely, when participants spent more time, their second-ranked ideas finished numerically—but not significantly—higher in novelty ($M = 4.32$, $SD = 0.74$) and usefulness ($M = 3.71$, $SD = 0.90$) as compared to first-ranked ideas in novelty ($M = 4.27$, $SD = 0.61$), $t(143) = 0.95$, $p = .35$, $d = 0.06$, and usefulness, ($M = 3.63$, $SD = 0.78$), $t(143) = 1.12$, $p = .27$, $d = 0.09$. Consistent with the aforementioned results on initial creativity, these results hint that to be deemed creative in Experiment 3, final ideas needed a balance of both novelty and usefulness, and that first-ranked initial ideas were more likely to achieve this balance with less time, whereas second-ranked initial ideas were more likely to achieve this balance with more time. The correlations in Table 2 suggest that raters in Experiment 3 weighted novelty ($r = 0.85$) and usefulness ($r = 0.85$) equally in their assessments of overall creativity, whereas raters in Experiment 2 weighted novelty ($r = 0.96$) more heavily than usefulness ($r = 0.64$). This is likely due to the nature of the tasks; raters may have seen usefulness as more important regarding automobile safety and novelty as more important in the relatively crowded domain of fitness equipment.

5.3.3. Initial abstractness

Consistent with the proposed theorizing, the initial ideas that participants ranked second tended to be more abstract than the initial ideas ranked first. Repeated-measures ANOVA showed that the three ranks significantly differed in initial abstractness, $F(2, 586) = 9.52$, $p < .001$, $\eta^2 = 0.03$. Post-hoc LSD tests showed that second-ranked initial ideas ($M = 3.99$, $SD = 0.91$) were significantly more abstract than first-ranked ideas ($M = 3.78$, $SD = 0.89$), $p < .001$, $d = 0.24$, but not significantly different from third-ranked ideas ($M = 3.95$, $SD = 0.88$), $p = .39$, $d = 0.04$. In contrast to Experiment 2, third-ranked ideas were significantly more abstract than first-ranked ideas, $p = .001$, $d = 0.21$. The more abstract third-ranked ideas in Experiment 3 were likely due to the procedures emphasizing abstractness, making all initial ideas more abstract on average. The key result in terms of the proposed theorizing is that second-ranked initial ideas were more abstract than first-ranked ideas.

[4] It is worth noting that there was a significant main effect for the order in which participants developed their top two initial ideas: the ideas they worked on first ($M = 3.94$, $SD = 0.70$) were significantly more creative in the end than the ideas they worked on second ($M = 3.68$, $SD = 0.76$), $t(293) = 6.28$, $p < .001$, $d = 0.37$. But results suggest that order impacted first-ranked and second-ranked initial ideas fairly equally, as the interaction between order and rank was not significant, nor was the three-way interaction between order, rank, and development time. However, including the second set of final ideas in the analyses reduced the effect sizes of the results, compared to the results using only the first set of final ideas.

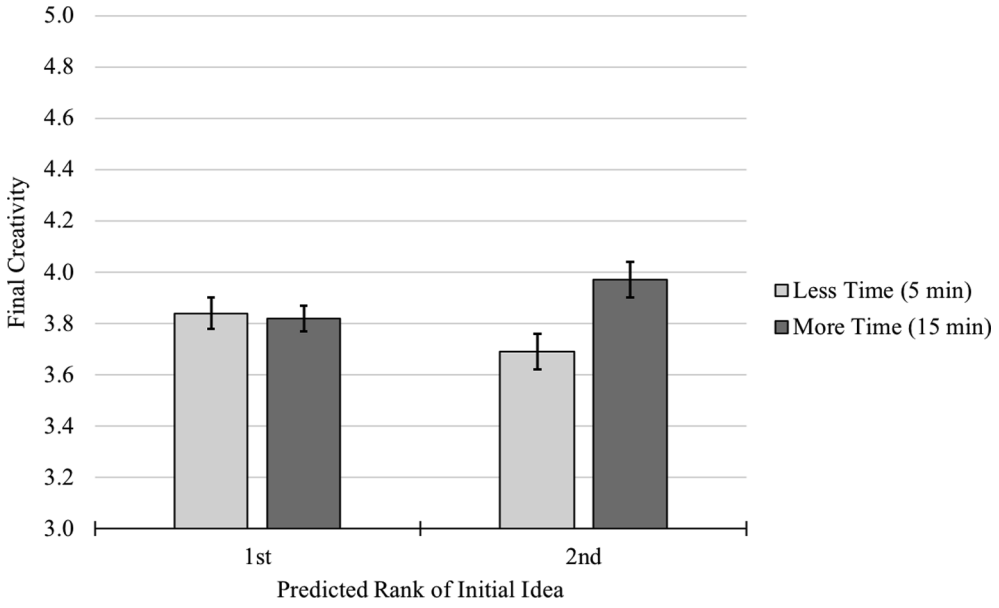

Fig. 5. Means from Experiment 3. Error bars are ± 1 SEM.

Table 6
Mixed-effects models for Experiment 3 testing whether abstract initial ideas improved more from development time than concrete initial ideas.

	Without Controlling for Initial Creativity			Controlling for Initial Creativity		
	df	*F*	*p*	*df*	*F*	*p*
Initial Abstractness	1	16.00	< 0.001	1	0.52	0.47
Time (less vs. more)	1	3.71	0.055	1	4.64	0.032
Initial Abstractness × Time	1	5.31	0.022	1	6.06	0.014
Initial Creativity (control)				1	182.95	< 0.001
	$F(3, 584) = 8.70$, $p < .001$			$F(4, 583) = 57.60$, $p < .001$		

Furthermore, mixed-effects models were used to test whether abstract initial ideas improved more from development time than concrete initial ideas (see Table 6). There was a significant interaction between initial abstractness and development time (less vs. more) without controlling for initial creativity, $F(1, 584) = 5.31$, $p = .022$, and when controlling for initial creativity, $F(1, 583) = 6.06$, $p = .014$. The nature of the interaction was similar to the pattern in Experiment 2 (see Fig. 4). When initial ideas were relatively abstract (one standard deviation above the mean in abstractness), more development time ($M = 3.96$, $SE = 0.06$) led to significantly greater increases in final creativity than less development time ($M = 3.71$, $SE = 0.06$), $t(583) = 3.25$, $p = .001$, $d = 0.27$. In contrast, when initial ideas were relatively concrete (one standard deviation below the mean in abstractness), more development time ($M = 3.79$, $SE = 0.06$) did not significantly change final creativity compared to less development time ($M = 3.80$, $SE = 0.06$), $t(583) = -0.11$, $p = .92$, $d = 0.01$. These results support the proposed theorizing, in that abstract initial ideas improved more from development time than concrete initial ideas.

6. Experiment 4: Addressing external validity and alternative explanations

6.1. Participants and procedures

The fourth experiment included 315 participants in the U.S. recruited via MTurk (44.13% female, age 20–72, $M_{age} = 37.42$, $SD_{age} = 11.72$), who were each compensated $6.00. Experiment 4 was designed to address the external validity of Experiments 1–3 and test key alternative explanations. The design built on Experiments 1–3 in four key ways.[5]

First, it is possible that ranking one's initial ideas on potential creativity may lack external validity. Perhaps the initial idea participants rank first in potential creativity is not the same initial idea they would pursue if left to their own devices. Experiment 4 addressed this possibility with five between-subjects conditions. Three conditions mirrored the prior experiments, in which participants ranked their three initial ideas on potential creativity, and then one of their three ideas was randomly selected for them to develop (labelled the random-first, random-second, and random-third conditions). In the fourth and fifth conditions, participants were allowed to pick which of their three initial ideas to develop. In the fourth condition, participants ranked their three initial ideas on potential creativity, but were told that they would be developing the idea that they ranked first (labelled the rank-pick condition). In the fifth condition, participants did not rank their three initial ideas, but instead, simply picked which one they wanted to develop (labelled the pick-only condition). Because ranking three initial ideas may involve more effort or deeper thought than simply picking

[5] To determine sample size, Experiment 3 was used to set a benchmark, as the manipulations best matched Experiment 4. Because development time was 15 min across all conditions in Experiment 4 and participants completed one final idea, the effect size between the 15-minute (more-time) conditions for the first final idea submitted in Experiment 3 was used as a benchmark ($f = 0.20$). Power analysis showed that a sample size of 305 would provide 80% power in comparing the five conditions in Experiment 4 (Faul et al., 2007). To ensure this target was met, 340 participants completed the survey, but 26 were omitted because they did not submit an actual final idea (each condition had 4–6 of the 26 errant responses). One participant was added to enable equally divisible rater groups, bringing the final sample to 315 participants.

one idea, the rank-pick condition was included to ensure a fair comparison to the three random conditions (Cooper & Richardson, 1986). Together, these five conditions helped test whether participants' predicted rankings reflected the initial idea that they would likely pursue when left to their own devices, as the ideas ranked first in the random conditions could be compared to the ideas chosen in the rank-pick and pick-only conditions.

Second, it is possible that individuals are more likely to identify their most promising initial idea if they are explicitly cued into long-term potential. Thus, in all five conditions, the instructions emphasized long-term potential. In the four conditions in which participants ranked their initial ideas, they were told: "Please rank your initial ideas based on their potential creativity in the long term, meaning how novel and useful they could become after you spend a long time developing them into finalized ideas." In the rank-pick condition, the instructions also said "…you will have a long time to further develop and ultimately finalize the idea you rank #1 below." In the pick-only condition, participants were told "…you will have a long time to further develop and ultimately finalize the idea you select below." In all five conditions, participants were then required to spend 15 min developing their respective initial ideas. Before that point, the instructions referred generally to a "long time" rather than the specific number of minutes, as specifying a timeframe may have undermined the emphasis on long-term potential.

Third, it is possible that spending more time evaluating their initial ideas may help individuals identify their most promising initial idea. Thus, in all five conditions, participants were told: "This step of the creative process is really important. So please spend at least five minutes thinking about your rankings [choice]." The survey would not advance until five minutes had passed.

Fourth, to complement Experiments 1–3, Experiment 4 focused on a different creative task. Participants were asked to develop a creative idea for a new travel experience to offer to consumers. All participants read the same introduction, which explained how consumers often seek novel travel experiences and treasure their memories of their favorite trips, and thus creative ideas in the travel industry can generate a lot of value for both consumers and the companies that offer them. Similar to Experiment 3, to foster more abstract initial ideas across conditions, participants were encouraged to combine elements of two concepts as they generated their three initial ideas (Berg, 2014). To build on Experiment 3, the concepts were captured in words rather than photos: one concept from the travel domain ("Train Tracks") and one that was relatively novel for the domain ("Improvisation"). The order of the two concepts was randomized. Participants were given the same incentive as Experiment 3: "If your final idea is in the top 10% of participants in overall creativity (novelty and usefulness), you will receive a $10 bonus."

6.2. Measures

6.2.1. Creativity

The overall creativity, novelty, and usefulness of participants' initial and final ideas were measured using the same procedures as Experiments 2 and 3. A separate sample of 300 consumers were recruited via MTurk to serve as raters (44.67% female, age 20–79, M_{age} = 37.16, SD_{age} = 11.53), who were each compensated $6.00. Expert ratings of the final ideas were also collected to complement the consumer ratings. Fifty travel experts were recruited and compensated via Qualtrics, including travel agents, hotel concierges, hotel managers, travel sales representatives, and tour guides (72.00% female, age 22–84, M_{age} = 41.14, SD_{age} = 14.34). They averaged 9.14 years of professional experience in the travel industry (SD = 9.09, range: 2–40).

Participants' 945 initial ideas were randomly divided into five groups of 189 ideas, and their 315 final ideas were randomly divided into five groups of 63 ideas. Each initial idea was rated by 30 consumers, and each final idea was rated by 30 consumers and ten experts (40 raters total). The consumer and expert ratings were highly correlated (r = 0.64 for creativity, $p < .001$), and results were similar when they were combined or analyzed separately, and thus the combined results were used (Appendix B reports the minor differences between consumers and experts). Each set of raters met conventional standards for interrater reliability—see Table 2 (LeBreton & Senter, 2008). An example of a final idea rated high in creativity was a cruise focused on volunteering and giving back to others (see Appendix A for full idea—the corresponding initial idea was also rated high in abstractness). An example of a final idea rated low in creativity was a scenic gondola ride in Italy (the corresponding initial idea, "gondola ride," was rated low in abstractness).

6.2.2. Initial abstractness

The abstractness of participants' 945 initial ideas was measured using the same procedures as Experiments 2 and 3 by an independent set of ten raters recruited from MTurk (70% female, age 27–66, M_{age} = 40.70, SD_{age} = 13.01), who were each paid $50.00.

6.3. Results and discussion

6.3.1. Initial abstractness

This experiment helped test the external validity of ranking initial ideas on potential creativity: is the initial idea that participants rank first the same initial idea they would decide to pursue on their own? The results detailed below revealed that in the two conditions where participants picked which initial idea to pursue, they chose ideas that were similarly concrete as the ideas ranked first by participants in the three random conditions, who submitted their rankings before being randomly assigned one of their initial ideas.

Because participants in the pick-only condition selected an initial idea to develop but did not rank their initial ideas, to test initial

Table 7
Means and SD's by condition for Experiment 4.

Condition	n	Initial Abstractness	Initial Creativity	Initial Novelty	Initial Usefulness	Final Creativity	Final Novelty	Final Usefulness
Random:								
1st Rank (Random-First)	63	4.30 (0.75)	3.89 (0.76)	3.93 (0.82)	3.87 (0.57)	4.02 (0.85)	4.16 (0.91)	4.07 (0.73)
2nd Rank (Random-Second)	62	4.48 (0.78)	3.70 (0.77)	3.83 (0.88)	3.79 (0.63)	4.47 (0.67)	4.50 (0.63)	4.33 (0.50)
3rd Rank (Random-Third)	62	4.41 (0.71)	3.57 (0.79)	3.68 (0.86)	3.68 (0.65)	4.22 (0.68)	4.36 (0.72)	4.19 (0.60)
Rank-Pick:								
1st Rank (Idea Selected)	64	4.17 (0.84)	3.74 (0.76)	3.84 (0.87)	3.89 (0.59)	4.13 (0.75)	4.30 (0.89)	4.17 (0.60)
2nd Rank		4.53 (0.76)	3.56 (0.69)	3.64 (0.77)	3.77 (0.58)			
3rd Rank		4.57 (0.77)	3.61 (0.68)	3.74 (0.79)	3.67 (0.58)			
Pick-Only:								
Idea Selected	64	4.22 (0.80)	3.72 (0.82)	3.82 (0.91)	3.83 (0.72)	4.11 (0.67)	4.22 (0.71)	4.16 (0.63)
Two Ideas Not Selected		4.41 (0.57)	3.53 (0.72)	3.64 (0.79)	3.68 (0.62)			

abstractness consistently across conditions, paired-samples *t*-tests were conducted between the initial idea that participants ranked first (or picked) and the average of their other two initial ideas (see means in Table 7). Initial ideas that participants ranked first (or picked) were significantly less abstract than the other two in the three random conditions, $t(186) = -2.44$, $p = .016$, $d = -0.16$, in the rank-pick condition, $t(63) = -3.68$, $p < .001$, $d = -0.45$, and in the pick-only condition, $t(63) = -2.10$, $p = .039$, $d = -0.27$. Moreover, the initial ideas that participants ranked first or picked did not significantly differ in abstractness between the random, rank-pick, and pick-only conditions, $F(2, 312) = 0.79$, $p = .46$, $\eta^2 = 0.005$, and no pairwise comparisons between conditions were significant. Thus, across all conditions, participants tended to favor concrete initial ideas over more abstract ones.

For the four conditions in which participants ranked their initial ideas, repeated-measures ANOVA was used to test whether the three predicted ranks differed in initial abstractness, and a between-subjects factor was included to test whether the rank-pick condition differed from the three random conditions. Results showed that the three predicted ranks significantly differed, $F(2, 498) = 10.19$, $p < .001$, $\eta^2 = 0.04$. Post-hoc LSD tests showed that first-ranked ideas were significantly less abstract than second-ranked ideas, $p < .001$, $d = -0.24$, and third-ranked ideas, $p < .001$, $d = -0.19$. Second- and third-ranked ideas did not significantly differ, $p = .77$, $d = 0.04$. The condition factor was marginal, $F(2, 498) = 2.47$, $p = .086$, $\eta^2 = 0.01$. The pattern of results was similar for the rank-pick and random conditions, but slightly more extreme in the rank-pick condition: first-ranked ideas were even less abstract than second- and third-ranked ideas (see effect sizes in prior paragraph). Thus, informing participants in the rank-pick condition that they would be developing their first-ranked idea did not significantly alter their rankings compared to the random conditions, in which participants did not know which idea they would develop.

To test whether abstract initial ideas improved more from development time than concrete initial ideas, a regression model was used, with initial abstractness predicting final creativity, controlling for initial creativity (see Table 8). Unlike the equivalent analysis in Experiments 2 and 3, there was no interaction term because all participants spent "more time" (15 min) developing their ideas. Initial abstractness was a significant predictor of final creativity, $\beta = 0.14$, $t = 2.92$, $p = .004$. Consistent with the prior experiments, relatively abstract initial ideas improved more from being developed than relatively concrete initial ideas.

6.3.2. Initial creativity

Consistent with H1, results showed that the initial ideas participants ranked first (or picked) were relatively creative in their initial form. In the three random conditions, first-ranked initial ideas were significantly higher than the other two ideas in creativity, $t(186) = 5.47$, $p < .001$, $d = 0.40$, novelty, $t(186) = 3.31$, $p = .001$, $d = 0.25$, and usefulness, $t(186) = 3.25$, $p = .001$, $d = 0.24$ (see means in Table 7). In the rank-pick condition, first-ranked initial ideas were significantly higher than the other two ideas in creativity, $t(63) = 2.12$, $p = .038$, $d = 0.26$, and usefulness, $t(63) = 2.79$, $p = .007$, $d = 0.34$, and marginally higher in novelty, $t(63) = 1.84$, $p = .071$, $d = 0.22$. In the pick-only condition, first-ranked initial ideas were significantly higher than the other two ideas in creativity, $t(63) = 2.43$, $p = .018$, $d = 0.31$, and usefulness, $t(63) = 2.39$, $p = .020$, $d = 0.32$, and marginally higher in novelty, $t(63) = 1.94$, $p = .058$, $d = 0.23$. Moreover, initial ideas ranked first or picked did not significantly differ between the random, rank-pick, and pick-only conditions in initial creativity, $F(2, 312) = 1.64$, $p = .20$, $\eta^2 = 0.01$, novelty, $F(2, 312) = 0.47$, $p = .62$, $\eta^2 = 0.003$, or usefulness, $F(2, 312) = 0.20$, $p = .82$, $\eta^2 = 0.001$, and no pairwise comparisons between conditions were significant. Thus, across all conditions, the initial ideas that participants ranked first or picked were relatively creative compared to their other initial ideas, but not significantly different from the initial ideas ranked first or picked in the other conditions.

For the four conditions in which participants ranked their initial ideas, repeated-measures ANOVA was used to test whether the three predicted ranks differed in initial creativity, and a between-subjects factor was included to test whether the rank-pick condition differed from the three random conditions. Results showed that the three ranks significantly differed in initial creativity, $F(2, 498) = 10.83$, $p < .001$, $\eta^2 = 0.04$, novelty, $F(2, 498) = 5.21$, $p = .006$, $\eta^2 = 0.02$, and usefulness, $F(2, 498) = 9.90$, $p < .001$, $\eta^2 = 0.04$. The condition factor was not significant for any of the dependent variables: creativity, $F(2, 498) = 2.30$, $p = .10$, $\eta^2 = 0.009$, novelty, $F(2, 498) = 2.18$, $p = .12$, $\eta^2 = 0.009$, or usefulness, $F(2, 498) = 0.16$, $p = .86$, $\eta^2 = 0.001$. This suggests that informing participants in the rank-pick condition that they would be developing their first-ranked idea did not significantly alter their rankings compared to the random conditions, in which participants did not know which idea they would develop.

Post-hoc LSD tests showed that participants' predicted rankings were largely accurate in terms of initial creativity. First-ranked ideas were significantly higher than second-ranked ideas in creativity, $p < .001$, $d = 0.28$, novelty, $p = .008$, $d = 0.17$, and usefulness, $p = .024$, $d = 0.13$. First-ranked ideas were also significantly higher than third-ranked ideas in creativity, $p < .001$, $d = 0.36$, novelty, $p = .005$, $d = 0.25$, and usefulness, $p < .001$, $d = 0.29$. These results support H1, as the initial idea participants ranked first tended to be their most creative. However, second-ranked ideas were not significantly higher than third-ranked ideas in creativity, $p = .42$, $d = 0.11$, or novelty, $p = .67$, $d = 0.10$, but were significantly higher in usefulness, $p = .026$, $d = 0.17$.

6.3.3. Final creativity

One-way ANOVA showed that the five conditions significantly differed in final creativity, $F(4, 310) = 3.52$, $p = .008$, $\eta^2 = 0.04$ (see means in Table 7). Consistent with H2, post-hoc LSD tests showed that ideas from the random-second condition finished significantly higher in creativity than random-first, $p = .001$, $d = 0.59$, rank-pick, $p = .008$, $d = 0.49$, and pick-only, $p = .005$, $d = 0.52$, and marginally higher than random-third, $p = .052$, $d = 0.37$.

Similar to Experiment 3, the differences in overall creativity appear to be driven by a combination of novelty and usefulness together. Random-second finished numerically higher than the other four conditions in both novelty and usefulness, but only some of these

Table 8
Regression models for Experiment 4 testing whether abstract initial ideas improved more from development time than concrete initial ideas.

	Without Controlling for Initial Creativity				Controlling for Initial Creativity			
	b	*SE*	*β*	*p*	*b*	*SE*	*β*	*p*
Initial Abstractness	0.03	0.05	0.03	0.63	0.13	0.05	0.14	0.004
Initial Creativity (control)					0.55	0.04	0.59	< 0.001
	$F(1, 313) = 0.23$, $p = .63$ $R^2 = 0.001$				$F(2, 312) = 79.53$, $p < .001$ $R^2 = 0.34$			

differences were significant. In terms of final novelty, random-second was significantly higher than random-first, $t(123) = 2.45$, $p = .016$, $d = 0.43$, and pick-only, $t(124) = 2.38$, $p = .019$, $d = 0.42$, but not random-third, $t(122) = 1.15$, $p = .25$, $d = 0.21$, or rank-pick, $t(124) = 1.50$, $p = .14$, $d = 0.26$. In terms of final usefulness, random-second was significantly higher than random-first, $t(123) = 2.31$, $p = .023$, $d = 0.42$, but not random-third, $t(122) = 1.37$, $p = .17$, $d = 0.25$, rank-pick, $t(124) = 1.62$, $p = .11$, $d = 0.29$, or pick-only, $t(124) = 1.63$, $p = .11$, $d = 0.30$.

In sum, the results from Experiment 4 support the external validity of ranking initial ideas on potential creativity. The initial ideas ranked first in the random conditions resembled the initial ideas chosen in the pick conditions in terms of initial abstractness and creativity (including how they compared to participants' other two ideas). This suggests that the initial ideas participants picked when left to their own devices were one in the same as the initial ideas they would rank first in potential creativity. Furthermore, Experiment 4 constructively replicated Experiments 1–3, in that second-ranked initial ideas finished the highest in creativity, although the advantage over third-ranked ideas was marginal in Experiment 4 (which makes sense given that third-ranked ideas were relatively abstract in Experiment 4).

7. Experiment 5: Overcoming myopia in forecasting potential creativity (H3)

Results from Experiments 1–4 suggest that individuals tend to be myopic in forecasting potential creativity, as they under-ranked their initial ideas with the greatest long-term potential in favor of initial ideas that reached their potential faster but ultimately finished lower in final creativity. These results raise the question of how this myopic tendency can be overcome, enabling individuals to identify their most promising initial idea. Insights from Construal Level Theory may provide a mechanism for helping individuals forecast the long-term potential of their initial ideas (Trope & Liberman, 2010; Wiesenfeld et al., 2017). Results from Experiments 2–4 suggest that participants undervalued abstractness—and overvalued concreteness—in ranking the potential creativity of their initial ideas. This suggests that participants were forecasting the potential of their initial ideas at a construal level that was lower (more concrete) than ideal. Their default construal level may be relatively functional for evaluating the current creativity of finalized ideas (Mueller et al., 2014), but dysfunctional for forecasting the potential creativity of initial ideas. Thinking at a higher (more abstract) construal level may help individuals see more potential in their relatively abstract initial ideas. In turn, they may avoid the mistake that participants often made in Experiments 2–4: ranking a concrete initial idea above the abstract idea that was actually their most promising. Thus, the following hypothesis is proposed, which is tested in Experiment 5:

Hypothesis 3. Forecasting the potential creativity of initial ideas at a higher construal level helps individuals identify their most promising initial idea.

7.1. Participants and procedures

Participants included 257 students and staff at a large university on the U.S. West Coast (34.63% female, age 18–67, $M_{age} = 26.30$, $SD_{age} = 8.93$), who were each compensated $18.00. To test H3, the design included three between-subjects conditions: high construal, low construal, and control. To build on the prior experiments, a different creative task was used. To foster psychological realism (Berkowitz & Donnerstein, 1982), the task focused on a domain familiar and important to participants: the focal university's athletics department. Participants were asked to develop creative ideas for improving ticket sales and attendance at athletic events for teams other than men's football and basketball. All participants read an introductory passage explaining that despite the fact that many of the university's teams had been highly successful in terms of wins and losses, many of their teams lagged behind other universities in attendance at home games. To foster realism and help ensure motivation was high across conditions, participants were informed that their final ideas would be passed along to the athletics department for their consideration. Participants were also given the same incentive as Experiments 3 and 4 (a $10 bonus if their final idea was in the top 10% in overall creativity).[6]

The procedures were similar to the prior experiments in that participants were asked to generate three initial ideas and rank them in terms of potential creativity. Similar to Experiments 3 and 4, to foster more abstract initial ideas across conditions, participants were encouraged to combine elements of two concepts as they generated their three initial ideas (Berg, 2014): one concept related to the domain of university athletics (a t-shirt with the university's name) and one from outside the domain (a graphing calculator). The concepts were presented in photos, and the order of the two concepts was randomized.

Development time and rank were not manipulated—all participants were informed that they would spend 15 min developing and finalizing their first-ranked initial idea. This was intended to isolate the construal level manipulations. The construal level manipulations were delivered after participants generated their three initial ideas, on the subsequent survey page, to ensure that all participants experienced the same procedures up to that point. To manipulate construal level, procedures were adapted from past experiments on construal level (Alter, Oppenheimer, & Zemla, 2010; Mueller et al., 2014). These procedures were based on the logic that thinking about *why* behaviors happen promotes higher construal levels, whereas thinking about *how* behaviors happen promotes lower construal levels (Trope & Liberman, 2010; Wiesenfeld et al., 2017). The high-construal condition was encouraged to think about *why* their rankings reflect potential creativity, and the low-construal condition was encouraged to think about *how* their thought process produced their rankings. In particular, participants in the high- and low-construal conditions read the following paragraph before they ranked their initial ideas (the bold text is what the high-construal condition saw, and the text in brackets is what the low-construal condition saw instead—this paragraph was not shown to the control condition):

As you figure out your rankings, please use the blank field below to explain **why** [how] you think your rankings reflect potential creativity, meaning how novel and useful each idea could become after you spend time developing it into a finalized idea. This is an opportunity for you to "think out loud" about **why your rankings reflect the potential creativity of your initial ideas** [how your thought process produces your rankings of the potential creativity of your initial ideas]. **Why** [How] did you end up ranking your ideas in your chosen order?

7.2. Measures

7.2.1. Creativity

The consensual assessment technique was again used to measure the overall creativity, novelty, and usefulness of participants' initial and final ideas (Amabile, 1996). The same definitions of the three dimensions and rating scale were used as Experiments 2–4. Three consumers (students from the focal university) rated the 771 initial ideas, and a separate set of three consumers/students rated the 257 final ideas. In

[6] To determine sample size, the average effect size from Experiments 2 and 3 between first- and second-ranked ideas in initial abstractness was used as a benchmark ($d = 0.22$), as detecting this difference was central to Experiment 5 and called for a larger sample than the test of final creativity. The equivalent effect size from Experiment 4 was not used because Experiment 4 was run after Experiment 5. Power analysis showed that 249 participants (83 per condition) would provide 80% power (Faul et al., 2007). To ensure this target was met, a total of 260 participants completed the survey, but three were omitted because they did not submit an actual final idea.

Table 9
Means and SD's by condition for Experiment 5.

	Condition		
	Why (n = 89)	How (n = 81)	Control (n = 87)
Initial Abstractness:			
1st Rank	4.34 (1.11)	3.87 (1.36)	3.78 (1.03)
2nd Rank	3.90 (1.18)	4.21 (1.10)	4.11 (1.16)
3rd Rank	3.87 (1.18)	4.13 (1.19)	4.16 (1.14)
Initial Creativity:			
1st Rank	2.48 (0.77)	2.66 (0.89)	2.83 (0.77)
2nd Rank	2.75 (0.91)	2.85 (0.82)	2.67 (0.92)
3rd Rank	2.71 (0.81)	2.69 (0.92)	2.71 (0.83)
Initial Novelty:			
1st Rank	2.54 (0.87)	2.77 (0.91)	3.00 (0.79)
2nd Rank	2.85 (1.00)	2.94 (0.88)	2.83 (1.02)
3rd Rank	2.75 (0.89)	2.81 (0.99)	2.95 (0.85)
Initial Usefulness:			
1st Rank	2.69 (0.80)	2.84 (0.84)	3.03 (0.76)
2nd Rank	2.99 (0.93)	3.00 (0.68)	2.95 (0.93)
3rd Rank	2.80 (0.78)	2.91 (0.88)	2.92 (0.79)
Final Creativity	3.40 (1.02)	2.97 (0.86)	3.06 (0.84)
Final Novelty	3.74 (1.18)	3.24 (1.18)	3.42 (1.23)
Final Usefulness	3.56 (0.71)	3.33 (0.71)	3.32 (0.74)

Note: In Experiment 5, participants developed their 1st-ranked initial ideas into their final ideas.

addition, two Marketing Directors from the university's athletics department served as expert raters for the final ideas. The consumer and expert ratings were highly correlated ($r = 0.56$ for creativity, $p < .001$), and the pattern of results was similar when they were combined or analyzed separately, and thus the combined results were used (Appendix B reports the differences between consumers and experts). Each set of raters met conventional standards for interrater reliability—see Table 2 (LeBreton & Senter, 2008). An example of a final idea rated high in creativity was distributing weekly stories about athletes' lives as students via social media (see Appendix A for full idea—the corresponding initial idea was rated high in abstractness). An example of a final idea rated low in creativity was giving away free t-shirts (many initial ideas focused on giving away free t-shirts, which tended to score low in abstractness).

7.2.2. Initial abstractness

An independent set of three students from the focal university rated the initial abstractness of the 771 initial ideas, using the same procedures as Experiments 2–4.

7.3. Results and discussion

7.3.1. Initial abstractness

See Table 9 for means and Fig. 6 for a visual of key results for Experiment 5. As aforementioned, participants were informed that they would be developing the initial idea that they ranked first in potential creativity. First-ranked ideas significantly differed in initial abstractness between the three conditions (high-construal, low-construal, control), $F(2, 254) = 5.73$, $p = .004$, $\eta^2 = 0.04$. As predicted, first-ranked ideas were significantly more abstract in the high-construal condition than the low-construal, $t(168) = 2.46$, $p = .015$, $d = 0.38$, and control conditions, $t(174) = 3.45$, $p = .001$, $d = 0.53$. The low-construal and control conditions did not significantly differ, $t(166) = 0.49$, $p = .62$, $d = 0.07$.

Within the high-construal condition, first-ranked ideas were significantly more abstract than second-ranked ideas, $t(88) = 2.64$, $p = .010$, $d = 0.28$, and third-ranked ideas, $t(88) = 2.83$, $p = .006$, $d = 0.30$, but second- and third-ranked ideas did not significantly differ, $t(88) = 0.17$, $p = .86$, $d = 0.02$. Within the low-construal condition, the three initial ideas did not significantly differ in abstractness, although first-ranked ideas were marginally less abstract than second-ranked ideas, $t(80) = -1.77$, $p = .080$, $d = -0.20$. Within the control condition, first-ranked ideas were marginally less abstract than second-ranked ideas, $t(86) = -1.92$, $p = .058$, $d = -0.20$, and significantly less abstract than third-ranked ideas, $t(86) = -2.44$, $p = .017$, $d = -0.26$, but second- and third-ranked ideas did not significantly differ $t(86) = -0.27$, $p = .79$, $d = -0.03$. As predicted, these results suggest that participants in the high-construal condition were more inclined to rank relatively abstract initial ideas first, whereas participants in the low-construal and control conditions tended to rank more concrete ideas first.

7.3.2. Initial creativity

First-ranked initial ideas significantly differed between the three conditions in overall creativity, $F(2, 254) = 4.08$, $p = .018$, $\eta^2 = 0.03$, novelty, $F(2, 254) = 6.24$, $p = .002$, $\eta^2 = 0.05$, and usefulness, $F(2, 254) = 4.02$, $p = .019$, $\eta^2 = 0.03$. Comparisons between each pair of conditions revealed that these results were driven mostly by differences between the high-construal and control conditions. First-ranked initial ideas in the high- and low-construal conditions did not significantly differ in initial creativity, novelty, or usefulness (see means in Table 9), although the high-construal condition was marginally lower than the low-construal condition in initial novelty, $t(168) = -1.69$, $p = .094$, $d = -0.26$. However, first-ranked initial ideas in the high-construal condition were significantly lower than the control condition in initial creativity, $t(174) = -3.00$, $p = .003$, $d = 0.45$, novelty, $t(174) = -3.63$, $p < .001$, $d = -0.55$, and usefulness, $t(174) = -2.89$, $p = .004$, $d = -0.44$. The low-construal and control conditions did not significantly differ in initial creativity, novelty, or usefulness, although the control condition was marginally greater than the low-construal condition in initial novelty, $t(166) = 1.73$, $p = .086$, $d = 0.27$. These results suggest that first-ranked ideas in the high-construal condition were less creative in their initial form than in the control condition, but that first-ranked ideas in the high- and low-construal conditions did not significantly differ in initial creativity.

7.3.3. Final creativity

Participants' final ideas significantly differed between the three conditions in overall creativity, $F(2, 254) = 5.25$, $p = .006$, $\eta^2 = 0.04$ (see means in Table 9 and Fig. 6). In support of H3, final ideas were significantly more creative in the high-construal condition than the low-construal condition, $t(168) = 2.94$, $p = .004$, $d = 0.46$, and the control condition, $t(174) = 2.38$, $p = .018$, $d = 0.36$. The low-construal and control conditions did not significantly differ, $t(166) = -0.70$, $p = .49$, $d = 0.11$. These results for overall creativity appear to be driven by a combination of both novelty and usefulness. The three conditions significantly differed in final novelty, $F(2, 254) = 3.85$, $p = .022$, $\eta^2 = 0.03$, and usefulness, $F(2, 254) = 3.17$, $p = .044$, $\eta^2 = 0.02$. The high-construal condition was significantly greater than the low-construal condition in both novelty, $t(168) = 2.77$, $p = .006$, $d = 0.42$, and usefulness, $t(168) = 2.13$, $p = .034$, $d = 0.32$. The high-construal condition was marginally greater than the control condition in novelty, $t(174) = 1.75$, $p = .081$, $d = 0.27$, and significantly greater in usefulness, $t(174) = 2.21$, $p = .029$, $d = 0.33$. The low-construal and control conditions did not significantly differ in novelty, $t(166) = -0.99$, $p = .33$, $d = 0.15$, or usefulness, $t(166) = 0.10$, $p = .92$, $d = 0.01$. In sum, the results from Experiment 5 support H3, in that forecasting the potential creativity of their initial ideas at a higher construal level helped participants identify their highest-potential

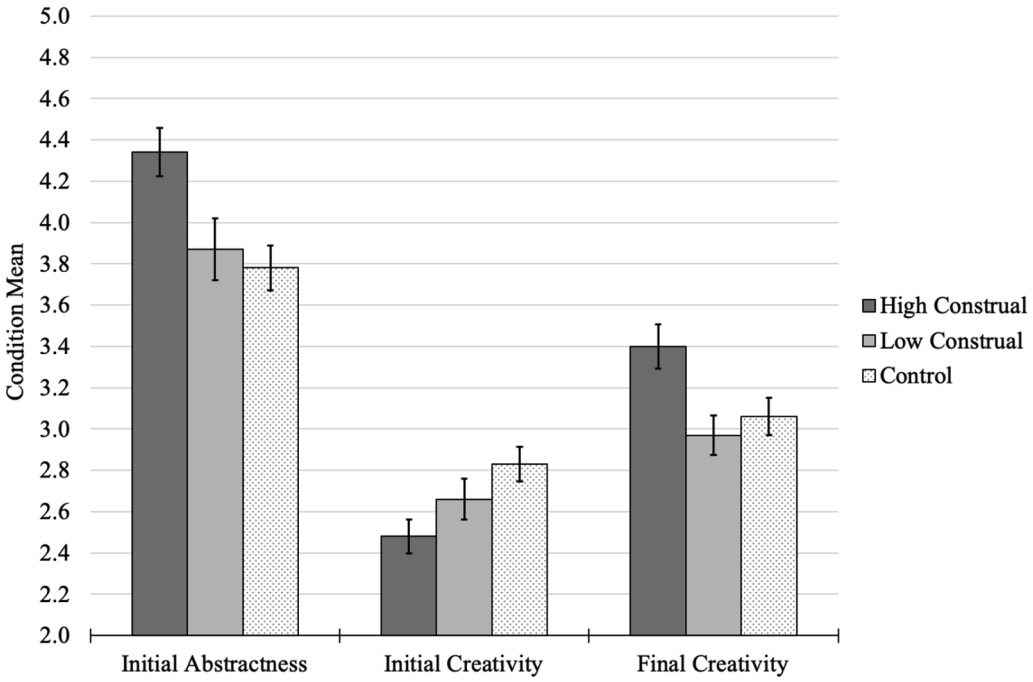

Fig. 6. Means by condition from Experiment 5. Initial abstractness/creativity capture the initial ideas participants ranked first, which they developed into their final ideas. Error bars are ± 1 SEM.

initial idea (which tended to be a relatively abstract initial idea).

Lastly, initial abstractness was tested as a moderated mediator between the high-construal condition (vs. low/control) and final creativity (Hayes, 2013). Development time was 15 min for all conditions, so the number of words participants generated in their "scratch paper" was used as a proxy for development, which did not significantly vary between conditions but had substantial variance among participants (M = 109.90 words, SD = 89.49). Given the proposed theorizing that abstract initial ideas improve more from development than concrete ideas, development was a second-stage moderator (Edwards & Lambert, 2007) and initial creativity was a control. As expected, the indirect effect was significant when development was high (one SD above the mean), [0.002, 0.12], and at the mean [0.004, 0.09], but not when development was low (one SD below the mean), [−0.006, 0.09]. This suggests that forecasting at a higher construal level led participants to select more abstract initial ideas, which translated into greater final creativity as long as they put effort into developing their ideas.

8. General discussion

Across five experiments, participants' accuracy in forecasting the potential creativity of their initial ideas was examined. Experiments 1–4 showed that participants were myopic in their forecasts of potential creativity: they undervalued abstract initial ideas with the greatest long-term potential and overvalued relatively concrete initial ideas with more obvious short-term potential. However, Experiment 5 showed that forecasting at a higher construal level helped participants overcome this myopic tendency and identify their most promising initial idea. These results have important implications for theory and research on creativity and construal level.

8.1. Theoretical implications

8.1.1. Idea evaluation and selection

Although the literature on creativity has traditionally focused on idea generation, a small but growing literature has emerged on idea evaluation and selection. This literature has largely focused on how individuals evaluate finalized ideas (e.g., Berg, 2016; Girotra et al., 2010; Mueller et al., 2012; Rietzschel et al., 2010; Runco & Basadur, 1993). The present research builds on this past work by shedding light on the related but distinct process of forecasting the potential creativity of initial ideas. The results suggest that forecasting potential creativity may present some unique challenges compared to evaluating the current creativity of finalized ideas. In particular, results across the five experiments suggest that individuals may be better at ranking their initial ideas on current creativity than potential creativity, highlighting the need to view these two processes separately. Much of the prior work on idea evaluation and selection suggests individuals are bad at assessing creative ideas (cf. Silvia, 2008). The present research suggests that in ranking initial ideas on potential creativity, accuracy depends on timing. Participants were actually fairly good at ranking their initial ideas in terms of short-term creativity, but struggled to forecast long-term potential. Broadly, this highlights the importance of timing in idea evaluation and selection. Only looking at one point in time may provide an incomplete or misleading picture of how idea evaluation and selection impact the creative process. These five experiments help illuminate the key role that idea evaluation and selection play early in the creative process, complementing prior work on these processes for finalized ideas.

8.1.2. BVSR model debate

This research helps integrate the two sides of the aforementioned debate over the Blind Variation and Selective Retention (BVSR) model of creativity. On one side, advocates of the BVSR model argue that individuals do not have any foresight about the potential creativity of their initial ideas (Campbell, 1960; Cziko, 1998; Simonton, 1999a, 1999b, 2003, 2011). On the other side, critics have argued that individuals can and do have some foresight about the potential creativity of their initial ideas (e.g., Gabora, 2007, 2011; Kozbelt, 2007, 2008; Mumford, 1999; Sternberg, 1998; Weisberg, 2004, 2015; Weisberg &

Hass, 2007).

The notion of myopic foresight uncovered in these experiments takes a step toward integrating these two perspectives. Results across the five experiments suggest that individuals have some degree of foresight about the potential creativity of their initial ideas, but that this foresight tends to be myopic in nature. Individuals may struggle to accurately forecast the long-term creativity of initial ideas when they are relatively rough and abstract. However, the finding that participants ranked their most promising initial idea second, and not third or fourth, suggests that they were better than chance in predicting creativity, even in the relative long term. Moreover, Experiment 5 suggests that forecasting at a higher construal level may help individuals recognize long-term potential earlier. Thus, despite individuals' myopic foresight, forecasting the potential creativity of initial ideas may not be a random walk as implied by the BVSR model. Instead, creators may be partially blind toward the beginning of the creative process, but more sighted if they use a higher construal level to evaluate their emerging ideas.

8.1.3. Construal level

This research also contributes to the growing literature on Construal Level Theory (Trope & Liberman, 2010). These experiments answer the call for more research on the notion of construal fit—i.e., the extent to which construal level matches the task at hand (Wiesenfeld et al., 2017). The results suggest that when it comes to forecasting the potential creativity of initial ideas, individuals' default construal level tends to be too low. This leads them to undervalue abstract initial ideas and overvalue concrete initial ideas. The results from Experiment 5 suggest that this problem can be mitigated to some extent, as encouraging participants to forecast potential creativity at a higher construal level helped them identify their most promising initial idea.

Although individuals' default construal level is often adaptive for the task at hand (Ledgerwood, Trope, & Liberman, 2010), the present research suggests that forecasting potential creativity may be a situation in which individuals' default construal level is maladaptive. This builds on prior research by Mueller et al. (2014), which hints that individuals' default construal level may be relatively adaptive for assessing the creativity of finalized ideas. Their results showed that dragging down participants' default construal level led to underestimates of creativity, but participants' default construal level started out appropriately high for estimating creativity. This is consistent with results from the present research regarding H1. Participants' default construal level was well-calibrated for evaluating the *current* creativity of their initial ideas before they changed much from their initial form. However, results for H2 and H3 suggest a higher construal level is helpful for forecasting the *potential* creativity of initial ideas that are bound to change before they are finalized. Taken together, these results suggest that the optimal construal level for evaluating ideas may be highest (most abstract) at the start of creative projects, and then become lower (more concrete) as ideas develop. In sum, the present research advances theory on construal level by shedding light on the important and potentially unique role it plays in forecasting the potential creativity of initial ideas.

8.2. Limitations and future directions

These five experiments have key limitations that may be addressed in future research. First, the creative tasks that participants tackled in the experiments were relatively brief compared to long-term creative projects in organizations, which may unfold over months or years. Although the relatively brief tasks allowed for controlled tests of the hypotheses, future research could explore forecasting potential creativity over longer time horizons. Allowing ideas to develop for more time could also enable more radical or breakthrough ideas to emerge. Second, across the experiments, participants generated three or four initial ideas. Although they tended to rank their highest-potential idea second, where the best idea falls in the rankings may be different when more initial ideas are at play. Future studies could examine the effects of generating more initial ideas before forecasting their potential. Third, across the experiments, participants who were assigned their third- or fourth-ranked ideas generated an approximately equal number of "scratch paper" words in elaborating their ideas as those assigned their first- or second-ranked ideas, hinting that participants were equally motivated regardless of the rank they were assigned. Nonetheless, it is still possible that motivation played a role in driving the results, which could be addressed in future experiments that manipulate motivation levels. Fourth, results in Experiment 2 seemed to be driven more by novelty than usefulness, whereas Experiments 3–5 seemed to be driven by a combination of novelty and usefulness together. This was likely due to differences in the tasks/domains and manipulations. Future research could unpack the role of novelty vs. usefulness in forecasting potential creativity, including how the two dimensions play out differently based on the task, domain, and other contextual factors. Fifth, participants in these experiments were non-expert creators. It is possible that expert creators have developed superior skills in forecasting potential creativity (Ericsson, 1999), or that expertise may undermine accuracy if the experts are too entrenched in their domains (Dane, 2010). Future work could compare the accuracy of expert and novice creators, which may reveal insights on the extent to which individuals can learn to accurately forecast potential creativity.

8.3. Practical implications and conclusion

This research may provide guidance to creators wishing to identify their most promising initial ideas. If pressed for time, it may be optimal to go with one's first instinct in selecting which initial idea to pursue, as this may be the quickest path to a moderately creative final idea. But if a highly creative idea is desired and time is not a concern, creators may benefit from deliberately resisting the allure of concrete initial ideas and pursuing one of their more abstract ideas instead. Broadly, the results suggest that creators may reject many of their best initial ideas without ever knowing it. To avoid this fate, creators may need to remember that their favorite initial idea is probably not their best idea.

Acknowledgments

I thank Scott Wiltermuth and anonymous reviewers for their helpful feedback. I thank the Stanford GSB Behavioral Lab for assistance with data collection. I am grateful to Richard Wright, Emily McLaughlin, Merrick Osborne, Nicole Abi-Esber, Aastha Chadha, Monica Nelson, Elizabeth Trinh, Zev Burstein, Sam Bernardon, Peter Schleede, Andrew Shoats, Ryan Cohen, Adit Ankur Desai, and Junwu Zhang for their generous help. I am also grateful to seminar attendees at the University of Washington and University of Southern California for their valuable comments. I thank the Wharton Mack Institute for Innovation Management for financial support.

Appendix A. Example ideas rated high in final creativity from Experiments 1–5

	Initial Idea	Notes during idea development	Final Idea
Experiment 1	VR mountain climbing	This would need a room of projected screens depicting a mountain scenery. The user would mount it like a treadmill, but it would be more like craggy stairs, like walking up the down escalator. The user would set out for a hike and walk up the stairs as the scenes depicted different natural elements- birds, waterfall, other wild animals, and flowers. There might be a wind machine. The strenuousness of the hike could vary with the slope of the escalator device. This device could provide an end goal that any exercise machines are missing. The goal in this case would be a scenic view	My idea is a VR hiking device. The user would mount it like a treadmill, but it would be more like craggy stairs, like walking up the down escalator. This would need a room of projected screens depicting a mountain scenery. The user would set out for a hike and walk up the stairs as the scenes depicted different natural elements—birds and other wild animals, waterfalls and flowers. The strenuousness of the hike could vary with the slope of the escalator device. There might be a wind machine as well. This device could provide an end goal that many exercise machines are missing. The goal in this case would be a beautiful scenic view
Experiment 2	My idea would be a gravity or wind machine. It would act like how you go into a pool with a current. You would be able to direct the gravity or wind to provide resistance to a certain direction. This would be free form resistance training without the wet mess of a pool	It would be difficult to create a machine that can create a wind current or gravity wherever you would like it to go. Due to this it can't just be a fan or even set of fans. Maybe you could actually do a magnet sort of deal. Have small magnets on your feet and hands attached like gloves or socks. Then you stand on a magnetic platform. The magnetic resistance would have to be adjustable of course. You could do that electronically. Maybe you could even reverse it to float a little bit. With magnets on your hands you could do handstands more easily and be able to do a variety of new exercises	A magnetic gravity machine. Gloves and shoes or socks with magnets on them combined with a magnetic platform. The platform magnetism can be adjusted according to the resistance you'd want. This would make a common thing like high-stepping a weight training exercise. You could also do handstand related exercise more easily. Reversing the magnetism you could levitate and do some exercises that involve your core muscles
Experiment 3	A slot machine is a type of game and a cupholder is small. Maybe you could make a small gamified interaction system - for example: using a small console, identify key things for the car to look out for (potholes, other speeding cars, etc.)	The console can have two gaming components: (1) Road observations: You can report things like potholes, construction, etc. (which means you must be watching the road). If the number of reports hits a certain threshold (ex: 50 people reporting), the alert appears in every autonomous vehicle. If you are one of the people who helped report the issue, you are given some kind of "prize" (ex: 10% discount to Dunkin Donuts or some other corporate partnership). (2) Route optimization: The console could have videogame like navigational arrows (up, down, right, left). The car will ask you to confirm each turn by clicking on the appropriate arrow. If you miss doing this a certain number of times (ex: 5), the car will automatically pull over and ask if you are okay or need medical assistance…	Each car will have an interactive console that requires interaction from the passenger and incentivizes participation. The console will have two functions: (1) Road observations: As a passenger in the car, you can report things like potholes, construction, etc. (which means you must be keeping your eyes on the road). If you report something, and a certain number of other people report the same issue (ex: 20 people on the road report the issue), the alert appears in every autonomous vehicle. Because you helped contribute to reporting the issue, you and everyone who reported the issue receives some kind of prize, such as a discount for coffee, discount on satellite radio, etc. (2) Route optimization: The console could have videogame-like navigational arrows (up, down, right, left). The car will ask you to confirm each turn the car is making by clicking on the appropriate arrow. If you miss more than 5 arrow notifications, the car will automatically pull over and ask if you are okay or need medical assistance.

Experiment 4	A cruise where participants can volunteer on-board to help other humans who have needs - so that participants can relax at sea but still feel good about themselves through volunteering	– Find ways to offer different type of volunteering projects on the cruise – Ensure that some of the people who will benefit are brought onboard the cruise – When the cruise arrives back in port, maybe the cruise participants then deliver the results of their volunteering work – Find ways to capture the benefits of the volunteering work in social media posts – Develop friendly competition on the cruise so that different teams can compete for recognition and achievement – Offer cruise discount or "refund stipend" if the cruise participants meet a certain goal with their volunteering efforts – Find ways to motivate the cruise participants to come up with new and innovative volunteering ideas as well (to build on the existing volunteering proposals) – Offer cruise discounts in the future if the participants meet a specific volunteering goal – Ensure that some of those who benefit from the volunteer work are able to speak to their appreciation, either by being onboard the cruise or by a recorded video that is shown to the cruise participants – Stop at various ports of call where the participants can participate in different volunteering projects on land in indigenous cultures – While at sea, conduct training on the different projects so that when the ship does stop at the various ports of call, the cruise participants are trained and more knowledgeable on their various roles and responsibilities for the volunteer projects (i.e., nominate a crew chief, clarify job tasks, train on use of specific tools or materials, etc.) – Bring experts onboard the cruise who can speak to the social, political or regional issues while at sea and why these volunteer efforts are so important for society – Ensure that drinks are available during these sessions so that participants can enjoy the experience	This unique travel experience is a cruise at sea with a theme of volunteering and giving back to others. While most cruises focus on self-indulgence, this type of cruise will appeal to those who want to give back to others and to help in an organized and structured activity. The cruise volunteer activities will occur both at sea and at specific ports of call. There will be a variety of volunteer activities that will appeal to the interests and skills of the cruise participants. In addition, while at sea, there will be various training and planning classes where cruise participants will learn how to use specific tools and will be trained on various activities, including maintenance, installation and repair. The cruise will also include various seminars and events where some of the people who benefit from this volunteer work will be able to show and express their appreciation to the cruise participants. One other benefit of this cruise is that if the cruise participants meet a specific volunteer goal, then there will be both a rebate on the price paid for this cruise as well as a substantial discount on future cruise opportunities. This opportunity will appeal to those people who want to give back to society but also want to have a fun and enjoyable experience at the same time
Experiment 5	More publicity on the student athletes as "students", so that other students will recognize their classmates and want to support them at games	Weekly spotlights on a different student athlete on the [university news] websites. Features on athletes on their academic department's page, talking about how they are involved in their major. Athlete written pieces about how they balance being a student and an athlete, how their academic areas of focus relate to their sport. Personal pitches from the athletes as to why students should attend their games, why their sport is cool, and what high crowd attendance means to them/how it improves their performance. A new athletics facebook/instagram/twitter account developed just for these student athlete stories and spotlights	One reason that more students do not attend certain sports teams' events is because they do not know anyone on the team, do not know about the sport or do not understand it very well. To combat this, the athletics and communications departments could work together to creates stories or features that focus on the student athletes both as students and as athletes. That way, students may recognize some of their peers in their classes or appreciate their accomplishments both inside and outside of the classroom better. They will be more likely to attend games to find out more about their classmates and to support them. The university could have weekly spotlights in a variety of campus news sources. Articles could also be posted on the athletes' academic departments' homepage or in their newsletter, that focus on why they chose their area of study and what they do as a student. They could also talk about how their academic interests relate to their athletic ones. These articles could be interviews or written by the athletes themselves. The athletes then could personally appeal to students and say why they think their sport is cool, why students should attend their games, and how good crowd attendance is meaningful to them/how it impacts their performance. The university could also create a new facebook/twitter/instagram account just for these stories, and they would probably reach an even larger amount of students who use these social media platforms

Appendix B. Differences in Results for Consumer vs. Expert Ratings

Weight Placed on Novelty vs. Usefulness in Overall Creativity (Experiments 2–5)

In comparing consumer and expert raters across the four experiments that had both (Experiments 2–5), consumers tended to place more weight on novelty and experts tended to place more weight on usefulness in their assessments of overall creativity. See the correlations in the table below. However, as reported in the main text, consumer and expert ratings were highly correlated in all four experiments, and the pattern of results was generally similar when each was analyzed separately (any differences are reported in the sections that follow). Thus, although consumers and experts displayed slightly different preferences in terms of novelty and usefulness, overall, they agreed more than they disagreed.

	Correlation with Overall Creativity (all $p < .001$)	
	Consumer	Expert
Experiment 2 (fitness equipment):		
Novelty	0.96	0.86
Usefulness	0.60	0.78
Experiment 3 (self-driving car safety):		
Novelty	0.87	0.61
Usefulness	0.83	0.90
Experiment 4 (travel experience):		
Novelty	0.95	0.84
Usefulness	0.82	0.85
Experiment 5 (university athletics marketing):		
Novelty	0.76	0.81
Usefulness	0.52	0.89

Experiment 2

The pattern of results was the same for consumer and expert ratings in terms of final creativity and final usefulness. The only difference concerns final novelty. For consumers, the six conditions significantly differed in final novelty, $F(5, 294) = 2.64$, $p = .024$, $\eta^2 = 0.04$, which mirrors the combined-ratings results reported in the main text. For experts, this result was marginal, $F(5, 294) = 2.08$, $p = .068$, $\eta^2 = 0.03$. For consumers, post-hoc LSD tests were the same pattern as the combined ratings: the second-more condition was significantly higher in final novelty than the other five conditions, and no other comparisons between the six conditions were significant. For experts, the second-more condition was numerically higher in final novelty than the other five conditions, and the differences were significant for second-less, $p = .021$, $d = 0.43$, third-less, $p = .003$, $d = 0.57$, and third-more, $p = .028$, $d = 0.42$, but not significant for first-less, $p = .084$, $d = 0.30$, and first-more, $p = .105$, $d = 0.32$.

Experiment 3

The pattern of results was the same for consumer and expert ratings in terms of final creativity and final novelty. But for final usefulness, the results were significant in the predicted direction when only the expert ratings were used, whereas the corresponding results were not significant for the consumer-only and combined ratings. Using only the expert ratings, when participants spent less time developing their top two initial ideas, first-ranked ideas ($M = 3.64$, $SD = 0.85$) finished significantly higher in usefulness than second-ranked ideas ($M = 3.45$, $SD = 0.91$), $t(149) = 2.55$, $p = .012$, $d = 0.21$. Conversely, when participants spent more time, second-ranked ideas ($M = 3.80$, $SD = 1.00$) finished significantly higher in usefulness than first-ranked ideas ($M = 3.63$, $SD = 0.84$), $t(143) = 2.28$, $p = .024$, $d = 0.19$.

Experiment 4

The pattern of results was the same for consumer and expert ratings in terms of final creativity, with two exceptions. First, experts rated the random-second condition ($M = 4.64$, $SD = 0.68$) significantly higher in final creativity than the random-third condition ($M = 4.21$, $SD = 0.69$), $t(122) = 3.49$, $p = .001$, $d = 0.63$. For consumers, random-second ($M = 4.42$, $SD = 0.75$) and random-third ($M = 4.22$, $SD = 0.75$) did not significantly differ, $t(122) = 1.47$, $p = .15$, $d = 0.27$. This difference was marginal for the combined ratings reported in the main text.

Second, in the regression model with initial abstractness predicting final creativity, controlling for initial creativity (see Table 7), the effect for initial abstractness was marginal using just the expert ratings, $\beta = 0.09$, $t = 1.66$, $p = .097$, but was significant with just the consumer ratings, $\beta = 0.14$, $t = 2.97$, $p = .003$. This effect was significant using the combined ratings, as reported in the main text.

There were also minor differences between consumers and experts in terms of final novelty and usefulness. In terms of final novelty, experts rated random-second ($M = 4.67$, $SD = 0.65$) higher than random-third ($M = 4.36$, $SD = 0.73$), $t(122) = 2.49$, $p = .014$, $d = 0.45$. For consumers, random-second ($M = 4.45$, $SD = 0.71$) and random-third ($M = 4.37$, $SD = 0.79$) did not significantly differ, $t(122) = 0.62$, $p = .54$, $d = 0.11$. The combined results reported in the main text matched the consumer results: random-second and random-third did not significantly differ.

In terms of final usefulness, experts rated random-second ($M = 4.44$, $SD = 0.66$) significantly higher than random-third ($M = 4.17$, $SD = 0.61$), $t(122) = 2.36$, $p = .020$, $d = 0.42$, and pick-only ($M = 4.20$, $SD = 0.70$), $t(124) = 2.04$, $p = .044$, $d = 0.35$. For consumers, random-second ($M = 4.29$, $SD = 0.54$) was not significantly higher than random-third ($M = 4.19$, $SD = 0.65$), $t(122) = 0.85$, $p = .40$, $d = 0.17$, or pick-only ($M = 4.15$, $SD = 0.68$), $t(124) = 1.25$, $p = .21$, $d = 0.23$. The combined results reported in the main text matched the consumer results: random-second did not significantly differ from random-third or pick-only.

Experiment 5

It is worth noting that the final ideas in Experiment 5 were rated by a total of three consumers and two experts, so each rater group was relatively small when analyzed separately. Nonetheless, the pattern of results was similar for consumers and experts, but the results were stronger for consumers than experts. For experts, the differences between the three conditions were marginal for creativity, $F(2, 254) = 2.44$, $p = .089$, $\eta^2 = 0.02$, and novelty, $F(2, 254) = 2.88$, $p = .058$, $\eta^2 = 0.02$, and were not significant for usefulness, $F(2, 254) = 1.25$, $p = .29$, $\eta^2 = 0.01$. These results were all significant for consumers and for the combined ratings reported in the main text.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451.

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*(5), 997–1013.

Amabile, T. M. (1996). *Creativity in context.* Boulder, CO: Westview Press.

Battelle, J. (2005). The birth of Google. Retrieved June 25, 2018, from https://www.wired.com/2005/08/battelle/.

Berg, J. M. (2014). The primal mark: How the beginning shapes the end in the development of creative ideas. *Organizational Behavior and Human Decision Processes, 125*(1), 1–17.

Berg, J. M. (2016). Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. *Administrative Science Quarterly, 61*(3), 433–468.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist, 37*(3), 245–257.

Blair, C. S., & Mumford, M. D. (2007). Errors in idea evaluation: Preference for the unoriginal? *Journal of Creative Behavior, 41*(3), 197–222.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological Review, 67*(6), 380–400.

Cooper, W. H., & Richardson, A. J. (1986). Unfair comparisons. *Journal of Applied Psychology, 71*(2), 179–184.

Cziko, G. A. (1998). From blind to creative: In defense of Donald Campbell's selectionist theory of human creativity. *Journal of Creative Behavior, 32*(3), 192–209.

Dahl, D. W., & Moreau, P. (2002). The influence and value of analogical thinking during new product ideation. *Journal of Marketing Research,* 47–60.

Dane, E. (2010). Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective. *Academy of Management Review, 35*(4), 579–603.

Diehl, M., & Stroebe, W. (1987). Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology, 53*(3), 497.

Diehl, M., & Stroebe, W. (1991). Productivity loss in idea-generating groups: Tracking down the blocking effect. *Journal of Personality and Social Psychology, 61*(3), 392.

Edwards, J., & Lambert, L. (2007). Methods for integrating moderation and mediation: A general analytical framework using moderated path analysis. *Psychological Methods, 12*(1), 1–22.

Ericsson, K. A. (1999). Creative expertise as superior reproducible performance: Innovative and flexible aspects of expert performance. *Psychological Inquiry, 10*(4), 329–333.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Finke, R. A. (1996). Imagery, creativity, and emergent structure. *Consciousness and Cognition, 5*(3), 381–393.

Förster, J., Friedman, R. S., & Liberman, N. (2004). Temporal construal effects on abstract and concrete thinking: Consequences for insight and creative cognition. *Journal of Personality and Social Psychology, 87*(2), 177–189.

Gabora, L. (2007). Why the creative process is not Darwinian: Comment on "The creative process in Picasso's Guernica sketches: Monotonic improvements versus non-monotonic variants". *Creativity Research Journal, 19*(4), 361–365.

Gabora, L. (2010). Revenge of the "neurds": Characterizing creative thought in terms of the structure and dynamics of memory. *Creativity Research Journal, 22*(1), 1–13.

Gabora, L. (2011). An analysis of the blind variation and selective retention theory of creativity. *Creativity Research Journal, 23*(2), 155–165.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou, & A. Ortony (Eds.). *Similarity and analogical reasoning* (pp. 199–241). Cambridge: Cambridge University Press.

Girotra, K., Terwiesch, C., & Ulrich, K. T. (2010). Idea generation and the quality of the best idea. *Management Science, 56*(4), 591–605.

Harrison, S. H., & Rouse, E. D. (2015). An inductive study of feedback interactions over the course of creative projects. *Academy of Management Journal, 58*(2), 375–404.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach.* New York: Guilford.

Kim, Y. J., & Zhong, C.-B. (2017). Ideas rise from chaos: Information structure and creativity. *Organizational Behavior and Human Decision Processes, 138*, 15–27.

Kornish, L. J., & Ulrich, K. T. (2014). The importance of the raw idea in innovation: Testing the sow's ear hypothesis. *Journal of Marketing Research, 51*(1), 14–26.

Kozbelt, A. (2007). A quantitative analysis of Beethoven as self-critic: Implications for psychological theories of musical creativity. *Psychology of Music, 35*(1), 144–168.

Kozbelt, A. (2008). Longitudinal hit ratios of classical composers: Reconciling "Darwinian" and expertise acquisition perspectives on lifespan creativity. *Psychology of Aesthetics, Creativity, and the Arts, 2*(4), 221–235.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852.

Ledgerwood, A., Trope, Y., & Liberman, N. (2010). Flexibility and consistency in evaluative responding: The function of construal level. In M. P. Zanna, & J. M. Olson (Vol. Eds.), *Advances in Experimental Social Psychology: Vol. 43*, (pp. 257–295). Academic Press.

Licuanan, B. F., Dailey, L. R., & Mumford, M. D. (2007). Idea evaluation: Error in evaluating highly original ideas. *Journal of Creative Behavior, 41*(1), 1–27.

Lubart, T. I. (2001). Models of the creative process: Past, present and future. *Creativity Research Journal, 13*(3/4), 295–308.

Lucas, B. J., & Nordgren, L. F. (2015). People underestimate the value of persistence for creative performance. *Journal of Personality and Social Psychology, 109*(2), 232–243.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502–517.

Mueller, J. S., Melwani, S., & Goncalo, J. A. (2012). The bias against creativity: Why people desire but reject creative ideas. *Psychological Science, 23*(1), 13–17.

Mueller, J. S., Wakslak, C. J., & Krishnan, V. (2014). Construing creativity: The how and why of recognizing creative ideas. *Journal of Experimental Social Psychology, 51*, 81–87.

Mumford, M. D. (1999). Blind variation or selective variation? Evaluative elements in creative thought. *Psychological Inquiry, 10*(4), 344–348.

Obstfeld, D. (2012). Creative projects: A less routine approach toward getting new things done. *Organization Science, 23*(6), 1571–1592.

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2010). The selection of creative ideas after individual idea generation: Choosing between creativity and impact. *British Journal of Psychology, 101*(1), 47–68.

Rowling, J. K. (2016). Retrieved June 25, 2018, from http://www.jkrowling.com/about/.

Runco, M. A., & Basadur, M. (1993). Assessing ideational and evaluative skills and creative styles and attitudes. *Creativity and Innovation Management, 2*(3), 166–173.

Runco, M., & Smith, W. (1992). Interpersonal and intrapersonal evaluations of creative ideas. *Personality and Individual Differences, 13*(3), 295–302.

Schwarz, N., Bless, H., Strack, F., Klumpp, G., Rittenauer-Schatka, H., & Simons, A. (1991). Ease of retrieval as information: Another look at the availability heuristic. *Journal of Personality and Social Psychology, 61*(2), 195.

Shalley, C. E. (1991). Effects of productivity goals, creativity goals, and personal discretion on individual creativity. *Journal of Applied Psychology, 76*(2), 179.

Silvia, P. J. (2008). Discernment and creativity: How well can people identify their most creative ideas? *Psychology of Aesthetics, Creativity, and the Arts, 2*(3), 139–146.

Simonton, D. K. (1999b). *Origins of genius: Darwinian perspectives on creativity.* New York: Oxford University Press.

Simonton, D. K. (1999a). Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychological Inquiry, 10*(4), 309–328.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: The integration of product, person, and process perspectives. *Psychological Bulletin, 129*(4), 475–494.

Simonton, D. K. (2007). The creative process in Picasso's Guernica Sketches: Monotonic improvements versus nonmonotonic variants. *Creativity Research Journal, 19*(4), 329–344.

Simonton, D. K. (2011). Creativity and discovery as blind variation: Campbell's (1960) BVSR model after the half-century mark. *Review of General Psychology, 15*(2), 158–174.

Simonton, D. K. (2012). Foresight, insight, oversight, and hindsight in scientific discovery: How sighted were Galileo's telescopic sightings? *Psychology of Aesthetics, Creativity, and the Arts, 6*(3), 243–254.

Simonton, D. K. (2015). Thomas Edison's creative career: The multilayered trajectory of trials, errors, failures, and triumphs. *Psychology of Aesthetics, Creativity, and the Arts, 9*(1), 2–14.

Smith, S. M. (1995). Fixation, incubation, and insight in memory and creative thinking. In S. M. Smith, T. M. Ward, & R. A. Finker (Eds.). *The creative cognition approach* (pp. 135–156). Cambridge, MA: MIT Press.

Sternberg, R. J. (1998). Cognitive mechanisms in human creativity: Is variation blind or sighted? *Journal of Creative Behavior, 32*(3), 159–176.

Sternberg, R. J., & Lubart, T. I. (1991). An investment theory of creativity and its development. *Human Development, 34*(1), 1–31.

Thompson, L., Gentner, D., & Loewenstein, J. (2000). Avoiding missed opportunities in managerial life: Analogical training more powerful than individual case training. *Organizational Behavior and Human Decision Processes, 82*(1), 60–75.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440–463.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General, 113*(2), 169–193.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232.

Ward, T. B. (1994). Structured imagination: The role of category structure in exemplar

generation. *Cognitive Psychology, 27*(1), 1–40.
Ward, T. B., Smith, S. M., & Finke, R. A. (1999). Creative cognition. In R. J. Sternberg (Ed.). *Handbook of creativity* (pp. 189–212). New York: Cambridge University Press.
Weisberg, R. W. (2004). On structure in the creative process: A quantitative case-study of the creation of Picasso's Guernica. *Empirical Studies of the Arts, 22*(1), 23–54.
Weisberg, R. W. (2015). Expertise, nonobvious creativity, and ordinary thinking in Edison and others: Integrating blindness and sightedness. *Psychology of Aesthetics, Creativity, and the Arts, 9*(1), 15–19.
Weisberg, R. W., & Hass, R. (2007). We are all partly right: Comment on Simonton. *Creativity Research Journal, 19*(4), 345–360.
Wiesenfeld, B. M., Reyt, J.-N., Brockner, J., & Trope, Y. (2017). Construal level theory in organizational research. *Annual Review of Organizational Psychology and Organizational Behavior, 4*(1), 367–400.