Creativity Contests: An Experimental Investigation of Eliciting Employee Creativity

Jasmijn Bol
Tulane University's A.B. Freeman School of Business. jbol@tulane.edu

Lisa LaViers
Tulane University's A.B. Freeman School of Business. llaviers@tulane.edu

Jason Sandvik
University of Arizona's Eller College of Management. sandvik@arizona.edu

Abstract

Running a contest can help managers elicit creative ideas from employees by providing employees with incentives to develop and share ideas that will help the firm. Little is known, however, about how contest design affects the outcomes of subjectively evaluated creativity-based contests. We conduct an experiment to investigate the impact of two contest design choices, the job role of the contest's evaluator and the number of prizes that participants compete for, on employee participation behavior. We also examine how these contest design choices impact the creativity of the submitted ideas. We find that using a peer of the employees as an evaluator increases the number of ideas shared, but it does not impact the number of unique participants who enter the contest. In addition, we find that using peer evaluators leads to an increase in the creativity of the ideas. We find that awarding more prizes to participants does not increase overall participation, but it does increase the number of ideas shared by employees from under-represented demographics. Awarding more prizes, however, reduces the creativity of the ideas. Together these results show that contest design choices have an important impact on employee creative idea-sharing and that managers should carefully consider how to tailor contests to fit their firms' needs.

Keywords: Creativity, Contests, Tournaments, Management Control System Design, Subjective Evaluation
JEL Classification: JEL Codes: J33, M40, M41, M52, O31

Accepted by Regina Wittenberg Moerman. We thank an anonymous reviewer, Raina Sayer, Robert Sayer, Nathan Seegert, Chris Stanton, and participants of the 2020 Utah Experimental Economics Conference, the 2020 JMAR Brownbag Series, the 2020 ENEAR Seminar Series, and seminar participants from the University of Toronto. We also thank Tulane University's A.B. Freeman School of Business and the Center for Growth and Opportunity for providing financial support for this project.

1. Introduction

Many employees have creative ideas about how to innovate within their organization because of their day-to-day involvement in the firm's operations. Creative ideas are those that have a combination of two different features: novelty, meaning they are new or previously unknown, and usefulness, meaning they provide a clear benefit to the firm [Amabile 1983, Runco and Jaeger 2012]. Unfortunately, employees often do not share their creative ideas because doing so requires time and effort [Morrison and Wilhelm Jr 2004, Garicano and Rayo 2017, Sandvik, Saouma, Seegert, and Stanton 2020]. Managers therefore need to design incentive structures that reward employees for idea-sharing. In this study, we focus on one type of incentive structure, a contest, which rewards a subset of employees for sharing their most creative ideas. Contests—specifically, innovation and creativity contests—are commonly seen in practice, such as the internal firm hack-a-thons used by companies like Facebook and Hasboro [Chang 2012, Walker 2019]. While such contests are common, much is still unknown about how changes in their design can affect participation decisions and the creativity of the submitted ideas [Adamczyk, Bullinger, and Möslein 2012]. This study contributes to the literature by examining the impact of two critical design choices—the contest evaluator's role in the firm and the number of prizes that participants compete for—on employees' idea-sharing. Insights into how to design creativity contests that facilitate idea-sharing are important because long-term firm value can significantly increase if organizations can harness employees' creative ideas.

The contest literature, which is also widely referred to in economics and accounting as the tournament literature, shows that contests do elicit increased effort from individuals [Lazear and Rosen 1981, Green and Stokey 1983]. Most of these studies consider settings in which a participant's total output increases with effort and can be objectively measured. For creative output, however, the relation between effort and output is not necessarily monotonic, and the output cannot be objectively measured [Amabile 1982, Unsworth 2001, Kachelmeier, Reichert, and Williamson 2008]. As a result, the findings in the literature do not readily extend to creativity contests in which

the outputs are evaluated subjectively. Considering the growing importance of creative efforts in modern businesses, it is important to extend the tournament literature by examining how contest design choices affect subjectively measured creative output.

In tournaments with objective evaluations, the evaluator's role in the firm is of no concern to participants because every evaluator should arrive at the same objective assessment of participants' output. In contests with subjective evaluation, however, we argue that the job role of the evaluator is of great importance to the participants because the winner is determined by the evaluator's taste function. For example, in a footrace with *subjective* evaluation, the winner might not be the fastest person but the one who looks the best while running (via an appealing running posture or an attractive outfit). As a result, the runner evaluated as the best-looking will depend on the evaluator's tastes. In this type of setting, before deciding to exert effort, participants will use perspective-taking to predict the evaluator's tastes and assess their likelihood of winning. Perspective-taking is an act in which a person tries to imagine the motivations and actions of another person [Batson, Early, and Salvarani 1997]. We posit that employees will believe that they can more accurately estimate the perspective, that is, the tastes, of the contest's evaluator when the evaluator has a role within the firm that is more similar to theirs [Parker and Axtell 2001, Epley, Keysar, Van Boven, and Gilovich 2004]. As a result, we predict that employee participation will be higher when the evaluator is a workplace peer, rather than a workplace manager, as such a situation will allow them to better estimate their probability of winning.

Also based on perspective-taking theory, we predict that contest participants will adjust the two creative features of their idea submissions to match what they perceive to be most important to the evaluator [Grant and Berry 2011, Grabner 2014]. Prior literature states that the creativity of an idea is measured based on an evaluator's subjective judgments of its novelty and usefulness within the context of the particular firm or situation [Amabile 1983]. Once an evaluator judges the novelty and usefulness of an idea, they weigh the relative importance of the two features and combine them into a single judgement of the creativity of the idea. We posit that employees will adjust their ideas' features, making the idea more useful or novel, depending on which of the two they believe

the evaluator will assign a larger weight to [Ostermaier and Uhl 2020]. We specifically predict that participants will submit ideas that are more useful when being evaluated by their managers because managers bear the costs of implementing the ideas in practice and are, therefore, likely to place more weight on idea usefulness when judging the creativity of the idea.

In addition to choosing the evaluators of the contest, the contest's organizers must also decide on the number and size of the prizes that participants compete for (i.e., the prize structure). In studies of tournaments with *objective* evaluation, the number of prizes offered has been shown to have little effect on participant behavior when the expected payoff is held constant [Cohen, Kaplan, and Sela 2008]. In contrast, we posit that the number of prizes competed for will affect participant behavior in a *subjectively* evaluated creativity contest. Subjective evaluation of creative output increases participants' uncertainty regarding their likelihood of winning. There is uncertainty both about the evaluator's tastes and about the effort that participants need to exert to improve their chances of winning. Due to this uncertainty, participants may feel that their likelihood of winning a subjectively evaluated winner-take-all creativity contest is not high enough to merit the provision of effort. We predict, however, that when competing for multiple prizes, even if the prizes are smaller, participants will feel like some of the uncertainty has been mitigated and that they are more likely to be rewarded. Instead of needing to submit the one idea that best fits the evaluator's tastes, their idea only needs to land in the top group. This is an easier threshold to clear, which will reduce the perceived uncertainty around winning, leading to increased participation. However, since this increased participation is largely driven by the submission of ideas that the participants believe to be less likely to win, we predict that when multiple small prizes are offered, instead of a large winner-take-all prize, the average creativity of the submitted ideas will decrease.

We conduct a two-stage experiment among online labor market participants to examine how participation in creativity contests is influenced by the role of the evaluator and the number of prizes participants compete for. We also examine how these contest design choices impact the creativity of the submitted ideas, as well as how they impact the two different features of the ideas, their novelty and usefulness. In the first stage of the experiment, participants complete a screening

survey. After that, those who indicate a willingness to return for a second task and who have extensive enough experience on the labor market platform are invited to participate in one of four creativity contests (i.e., the second stage). Participants are randomized into contests following a two-by-two design. In the first treatment, the role of the evaluator is manipulated by assigning a panel of online labor market participants (i.e., the peers of the contest participants) as the evaluators of the ideas for half of the participants, while a panel of individuals who regularly conduct surveys among online labor market participants (i.e., the managers of the contest participants) are assigned as evaluators for the other half. The second treatment varies the number of prizes while holding fixed the total amount of prize money. We manipulate this by having half of the participants compete for one $100 prize and the other half compete for ten $10 prizes.

The contests vary in these two design choices, but the task is the same across all four contests: we ask participants to submit their most creative ideas for a survey attention check question that could be used to engender better quality data from future survey-takers. Participants are informed of the role of their contest's evaluator and their contest's prize structure, and then they are asked to submit their ideas for attention checks. Due to the two-stage nature of our design, we can study the impact of contest design on *both* participation behavior and the creativity of the submitted ideas. Our setting creates an environment similar to an actual workplace, where employees are asked, but not forced, to submit ideas in a creativity contest. Before submitting an idea, employees can assess whether their likelihood of winning is sufficient to merit their effort, allowing us to measure participation responses. We also asked participants to submit ideas related to their job, so they were not required to think about something entirely new to them. Our study, therefore, focuses on the sharing of job-related creative ideas in a real-world setting, making it a framed field experiment [Harrison and List 2004, Bloomfield, Nelson, and Soltes 2016].

In line with our predictions, we find that when participants are asked to submit their ideas in a subjectively evaluated creativity contest, the job role of the evaluator and the prize structure significantly affect their behaviors. We consider the effects of contest design choices on two different measures of participation: a discrete measure of the number of ideas submitted by each participant

and a binary measure of whether participants choose to enter the contest by submitting at least one idea. We find that participants submit more ideas when the evaluators are their peers, rather than their managers, suggesting that participants feel they have a better chance of winning when their ideas are evaluated by someone with their same job role. When we consider a participant's decision to submit at least one idea, we find a positive, though insignificant, effect of peer evaluators. These findings suggest that using peer evaluators is an effective way to increase the number of ideas collected but not the total number of employees who decide to participate.

When we examine the effect of prize structure on participation, we do not find a significant effect nor do we find a significant interaction between prize structure and the role of the evaluator on participation in the full sample. In order to explore this relation further, we examine the subset of our participants that provides a stronger test of the theory behind our predictions. Specifically, we examine the behavior of participants who are demographically under-represented in creative endeavors. Since young males are more likely to be seen as innovators [Hofstra, Kulkarni, Galvez, He, Jurafsky, and McFarland 2020], participants who are older and/or non-male will likely have lower assessments of their own win-likelihood in creativity contests, so multiple-prize incentives should have a stronger effect on their participation behavior. In support of this notion, we find that older and/or non-male individuals submit significantly more ideas when they compete for multiple small prizes, rather than a single large prize, whereas prize structure has an insignificant effect on the number of ideas submitted by young male individuals. Again, we do not find significant effects of prize structure on the decision to submit at least one idea, which further suggests that the studied contest design choices impact the total number of ideas collected but not the total number of individuals who participate.

To identify the effects of the contest design choices on idea creativity, we estimate the effects on both the average creativity across all submitted ideas and on the creativity of only the best submissions. We measure the two features of creativity using the evaluators' subjectively assigned ratings of the novelty and usefulness of each idea. Then, we measure idea creativity using the evaluators' subjectively assigned ratings of the creativity of each idea. When considering average

effects across all ideas, we find that participants submit ideas that are significantly less (more) useful when being evaluated by their peers (managers). This evidence is consistent with our prediction that participants adjust the two features of their ideas depending on what they perceive the focus will be of their assigned evaluator. As for the impact of the evaluator's role on the usefulness of the best ideas, we do not find a significant effect. However, we do find that the creativity of the best ideas significantly increases when peer evaluators are used, and we find that this increase is driven by increased idea novelty.

As for prize structure, we find that changing the number of prizes that participants compete for does not impact the average usefulness of the submitted ideas, nor does it impact the usefulness of the best ideas. We do find that the average novelty of the ideas decreases when multiple small prizes are offered, instead of a single winner-take-all prize. This leads to a decrease in the average creativity of the ideas. The adverse effect on novelty is magnified when we focus on the best submissions, suggesting that participants exert less effort to submit highly novel ideas when they compete for multiple small prizes. We provide further evidence in support of this conclusion through the use of multiple supplementary proxies of effort exertion. In particular, we show that participants reduce the amount of time they spend developing their ideas and they submit significantly shorter ideas when competing for multiple small prizes, rather than one large prize. As with participation behavior, we find little to no evidence of a significant interaction effect of the evaluator's role and the prize structure on the creativity of the submitted ideas.

Our study contributes to the literature on creativity and innovation [Amabile 1983, Adamczyk et al. 2012, Amabile 2012]. This literature has mainly focused on idea origination and, in general, shows that it is hard to incentivize the generation of creative ideas [Kachelmeier et al. 2008, Gneezy, Meier, and Rey-Biel 2011]. Kachelmeier, Wang, and Williamson [2019] find that the effects of incentives on creativity vary based on whether or not participants have time to incubate their ideas. Our study builds on their finding by employing field-like methods that ask participants to submit ideas that they may have already had throughout the course of their careers. Our findings suggest that although it might be difficult to incentivize "the light bulb moment" with traditional

incentive systems, this conclusion does not extend to the further development and dissemination of creative ideas. Our results show that changes to the design of the contest can impact both employees' participation behaviors and the creativity of employees' submitted ideas, therefore showing the manageability of this step through incentive structures.

In addition, we build on an important nuance in the creativity literature that distinguishes between the impact that incentives have on the creativity of the average idea and the best ideas [Kachelmeier et al. 2008, Kachelmeier and Williamson 2010]. Our study indicates that this nuance continues to be important for managers, as contest design choices do not always impact the submission of average and best ideas similarly. Management should therefore customize their creativity contest design choices to align with their goals. That is, our results suggest that the design should be different when, for example, the objective of the contest is identifying a single ground-breaking idea versus when management is seeking a wide array of useful ideas to implement continuous improvements throughout the firm.

Our study also contributes to the tournament literature by highlighting the important implications of subjective versus objective assessments of performance [Green and Stokey 1983, Connelly, Tihanyi, Crook, and Gangloff 2014, Dechenaux, Kovenock, and Sheremeta 2015]. Prior literature has not considered changes in the job role of the evaluator because in tournaments with *objective* performance evaluation, the evaluator's role does not impact their evaluations [Terwiesch and Xu 2008, Jeppesen and Lakhani 2010, Boudreau, Lacetera, and Lakhani 2011]. In a similar vein, prior research finds that the prize structure of an objectively evaluated contest does not have a strong effect on behavior, as long as the expected value of participating is held constant [Moldovanu and Sela 2001, Cohen et al. 2008]. We show, however, that these results do not extend to tournaments where performance is assessed subjectively. Subjective assessment increases uncertainty, reducing participants' ability to estimate their own chances of winning. We find that contest designers can alleviate some of this uncertainty by designating someone similar to the participants to evaluate the submissions. Our results also suggest that using multiple smaller prizes can reduce win-likelihood uncertainty, but only for participants who are demographically under-represented in creative en-

deavors. This finding may be particularly important for firms that are trying to be more inclusive by eliciting innovative ideas from groups of employees who have traditionally been under-represented when it comes to creative idea-sharing.

2. Motivation and Hypothesis Development

Firm managers want to elicit employees' creative ideas because they are the "fuzzy front end" of innovation, the process by which firms turn creative ideas into increased profits [Reid and De Brentani 2004, p.171]. For firms to elicit employees' creative ideas, two things must happen: (1) the employees must have initial ideas, that is, a "light bulb moment," and (2) the employees must develop their ideas and share them with management [Van Dijk and Van Den Ende 2002]. In this study, we focus on the second step, the development and sharing of creative ideas. We assume that some portion of employees will naturally think of creative ideas that can potentially benefit the firm during their day-to-day jobs, but without the proper incentive systems, these employees may not further develop and share their ideas because doing so is costly [Fairbank and Williams 2001, Fairbank, Spangler, and Williams 2003]. Organizations should, therefore, implement systems that motivate employees to develop and share their creative ideas [Menzel 2017].

There are a number of potential ways in which organizations can motivate creative idea-sharing, for example, through implicit rewards, like promotions, or explicit ones, like paying employees for every idea submitted. While each system has merits, in this study, we focus on contests as a means of eliciting creative idea-sharing. Contests can be effective motivators. Employees will exert effort based on their belief that they may be rewarded with a prize. Moreover, contests direct effort by providing a clear set of rules and instructions concerning the timing and format by which ideas should be shared. The motivating and directing elements of contests have made them popular methods for collecting creative ideas for hundreds of years, from Napoleon's contests—which yielded important advances in military readiness, like canning and margarine, which make food easier to transport—to Facebook's famous annual hack-a-thons—which are cited as one of the most important innovation drivers of the firm, leading to features like chat and calendars [Adamczyk et al.

2012, Chang 2012]. Importantly, nearly every company can stage a creative idea-sharing contest to complement its existing compensation structure without overhauling that system [Bradler, Neckermann, and Warnke 2019].

Contest design choices have been broadly studied in the accounting, management, and economics literatures in the context of motivating effort [Lazear and Rosen 1981, Connelly et al. 2014]. In most of these studies, performance monotonically increases with effort, and performance outcomes are objectively measured. This is frequently the case even when studying innovation. For example, participants may be asked to solve an unsolved math problem [Kiersz 2008]. While it takes creativity to solve the problem, contest evaluators can objectively tell which solutions solve the problem. As another example of a research study of innovation, Dean and Image [2008] examine participants who have been asked to share ideas about how to clean up oil spills in Arctic areas, where the winner is the individual who manages to prevent oil from freezing while being collected. This task requires creativity, but the output of the solution is objectively evaluated based on efficacy rates and cost. In our study, however, we focus on the common situation where the value of a proposed idea or solution cannot be objectively measured, but instead needs to be subjectively evaluated.

When the value of a creative idea is assessed subjectively by an evaluator, it matters who the evaluator is, as individuals all have unique tastes and will potentially value creative ideas differently. Subjective evaluation, therefore, creates additional uncertainty for employees who are invited to share their ideas in a contest. Not only is there uncertainty about the amount of effort needed to win when performance does not necessarily monotonically increase with effort, but there is also uncertainty about how the employee's output will be valued. This increased uncertainty influences the perceived trade-offs of entering the contest. As such, it is important to study creativity contest design choices that influence perceived uncertainty and, ultimately, participation decisions and the creativity of the ideas submitted by participants. In this research, we focus on two important contest design choices: the job role of the evaluator and the number of prizes that participants compete for. While there may be other critical design elements, we focus on these two because

they capture two fundamental choices that managers must make when holding a contest.

2.1. EVALUATOR ROLE

In a contest with subjective evaluations, someone or a group of people must be designated to evaluate the submissions. We argue that the job role of the evaluator affects employee participation in creative idea-sharing contests. We specifically examine the impact of the two most readily accessible and common roles of evaluators: (1) the employees at the firm, that is, the peers of the participants, who have relatively similar jobs, and (2) the managers of the firm, that is, the bosses of the participants, who have relatively dissimilar jobs. We posit that the employees invited to participate in the contest will use perspective-taking to predict the tastes of the designated evaluator of the creative ideas [Mead 1934, Ostermaier and Uhl 2020]. Several studies show that individuals feel they better understand the perspectives of those they resemble, which increases their confidence in their estimations of how those individuals will behave [Parker and Axtell 2001, Epley et al. 2004]. Hence, since employees and their workplace peers have similar day-to-day experiences (e.g., they complete similar tasks, have similar incentive schemes, and experience similar difficulties), whereas employees and their managers have less similar day-to-day experiences, employees will likely believe it to be easier to imagine what their peers will consider creative. Consequently, we predict that the evaluator's role in the firm will influence participation behavior because the evaluator's role influences the certainty with which employees believe they can estimate their chances of winning. Specifically, we predict that, in a subjectively evaluated creativity contest, using peer evaluators rather than manager evaluators will increase participation.

We also argue that the role of the evaluator will affect the creative features of the submitted ideas. As discussed previously, we follow prior literature by considering both the creativity of the ideas, as well as the two different features that make an idea creative, its novelty and usefulness [Amabile 1983, Runco and Jaeger 2012].[1] Once an evaluator judges the novelty and usefulness of an idea, they weigh the relative importance of the two features and combine them into a single

[1] In all our tests of idea creativity, we separately consider the impact of the contest design choices on the creativity, novelty, and usefulness of the ideas.

judgement of the creativity of the idea. We do not, however, expect that every evaluator puts the same weights on these two features of creativity. We predict that manager evaluators—who are in charge of implementing and funding these ideas to increase firm value and who are held accountable for the profits gained or lost—will emphasize the usefulness of the idea more than its novelty. Peer evaluators, on the other hand, are not likely to be in charge of implementing and funding the ideas and thus are likely to be less concerned about usefulness. We predict that employees will, through perspective-taking, anticipate this difference in focus between evaluators and submit ideas that are more (less) useful when evaluated by their managers (peers). However, it is unclear whether or not the increased usefulness of the ideas submitted to managers will also increase the creativity of the ideas or if the increased usefulness will trade off with decreased novelty, resulting in ideas that are less creative.[2]

Our prediction that the evaluator's role will impact idea usefulness speaks to shifting the distribution of ideas to the right, as all contributors try to make their ideas more appealing to the assigned evaluator. Such a distributional shift would lead to an increase in average idea usefulness. Most managers, however, are mainly focused on the best ideas, and consequently they are mainly concerned about whether the evaluator's role impacts the creativity of the best ideas. It is possible that contributors of the best ideas already strive to make their proposals as creative as possible, regardless of the role of the evaluator. If this is true, the creativity of the best ideas should not be impacted by the choice of evaluator. If, however, contributors do trade off usefulness and novelty depending on the evaluator's role, or if there is capacity for increased usefulness and novelty even among the best ideas, we may see an effect when focusing on only these ideas. Thus, considering the importance of the best ideas to managers, we pose a research question specifically aimed at examining the impact of the contest design choices on the creativity of the best ideas.

[2]Novelty and usefulness are not necessarily mutually exclusive features of creativity. The most creative ideas are those that are both highly novel and highly useful.

2.2. PRIZE STRUCTURE

The prize structure of contests can vary greatly, as contests range from winner-take-all models to participation-award models. In a winner-take-all contest, a single participant wins the entire prize pool. With participation awards, everyone receives a (usually small) reward for participating. Research on contests with objectively determined winners finds that, in a contest for which potential participants must expend effort to participate, only the expected payoff affects their behavior [Cohen et al. 2008].[3] This analytical result depends on allowing participants to select into and out of the contest, where contests with more valuable rewards will attract more participants.[4] As a result, variations in the number of prizes offered has little effect on behavior, so long as the expected value of participating is held constant. However, we predict that in a *subjectively* evaluated contest, the number of prizes that participants compete for will change behavior.

In a contest where higher effort directly leads to higher performance, only top performers are likely to believe they have a higher likelihood of winning a prize when the number of prizes increases. This is because, when evaluation is objective, participants can accurately assess what their output is likely to be, given their level of exertion and ability. Low performers may know that, even though the number of winners has increased, their output will still likely fall below the threshold for a prize, so the increase in the number of prizes will not influence their participation decisions. In a contest with subjective evaluation, however, where uncertainty about the evaluator's tastes obscures who the top performers might be, increasing the number of prizes will likely positively influence the participants' estimated win-likelihood, as their ideas do not need to perfectly match the evaluator's unknown taste function. Instead, their ideas just need to belong to the top group of ideas that are being rewarded. That is, if there are multiple prizes awarded, instead of a winner-

[3]The type of contest in Cohen et al. [2008] is also known as an all-pay contest in which all effort is expended before the winner is announced. Examples of all-pay contests are races, art shows, and basketball games. This type of contest is different than, for example, a contest for the best building design, where drafts and digital renderings of potential buildings are submitted, not completed buildings [Siegel 2009].

[4]The contest literature is quite large, and many studies examine optimal prize structures. One of the most famous of these, Moldovanu and Sela [2001], finds that the optimal number of prizes can be one or more than one, depending on whether participants' cost functions are linear, concave, or convex. Their result is particularly difficult to translate to the setting of creativity, where effort does not have such a direct relationship with output, making the cost function impossible to estimate.

take-all format, a larger group of employees will feel the expected value of participating is high enough for them to exert effort. As a result, we predict that, in a subjectively evaluated creativity contest, using multiple small prizes, rather than one large prize, will increase participation.

Moreover, we argue not only that the number of prizes affects participation but also that there is a trade-off between the number of ideas shared and the average creativity of those ideas. When there is only one large reward, participants will evaluate their chances of winning as lower. Employees who believe that their ideas are potentially creative enough to win the single, large reward will exert increased effort to develop and submit their ideas to win. However, employees who feel that their idea is not creative enough will likely choose not to participate, resulting in fewer low creativity ideas getting submitted to the contest. Because of this potential trade-off between the number of ideas submitted and the creativity of those ideas, offering a high number of small prizes, which invites increased participation from individuals with only marginally creative ideas, will decrease the average creativity of the submitted ideas. That is, we predict that, in a subjectively evaluated creativity contest, having participants compete for multiple prizes, rather than for one large prize, will decrease the average creativity of the ideas.

As before, we are not only interested in the average creativity of the submitted ideas, but also in the creativity of the best ideas. It is not clear what the effect of multiple prizes will be on the best ideas. They may not be affected, as theory predicts that increased participation will be driven by participants who believe they have a lower win-likelihood. Participants who estimate their chances of winning to be high might, therefore, be unaffected by the number of prizes, as they were going to submit their ideas no matter how many prizes were awarded. In this case, the creativity of the best ideas would be unaffected by incentive prize structure. On the other hand, the use of multiple small prizes, rather than a large winner-take-all reward, may actually reduce the creativity of the best ideas. Participants may not be as motivated by the low value of the smaller prizes as they are by the high value of the winner-take-all prize. As a result, they may exert less effort into their idea submissions. This reduced effort exertion would result in less creative ideas, even among the best ideas. As before, due to the lack of a clear directional prediction we state a research question about

the impact of contest design choices on the creativity of the best ideas.

2.3. INTERACTION EFFECTS

The two constructs that we examine in this study, the evaluator's role and the incentive prize structure, are chosen because they are both key elements of a contest and, practically speaking, both are design choices that need to be made to actually run a contest. They were not selected because of a theory-based belief that they must interact with one another. That being said, the two constructs may amplify each other if, for example, the perspective-taking benefits of peer evaluators only emerge when multiple prizes are offered. On the other hand, the two constructs may cancel each other out if, for example, participants perceive a contest with peer evaluators and many small prizes to be too informal and unrewarding to merit their effort. Alternatively, the two design choices might not interact at all. Because of the practical relevance of answering this question but the lack of theoretical foundation, we form additional research questions to examine the interaction effect of the two design choices on participation behavior and idea creativity. In the table below, we organize and formalize our hypotheses and research questions.

Hypotheses and Research Questions

Participation Behavior	
Hypothesis 1	In a subjectively evaluated creativity contest, using peer evaluators, rather than manager evaluators, will increase participation.
Hypothesis 2	In a subjectively evaluated creativity contest, having participants compete for multiple small prizes, rather than one large prize, will increase participation.
Idea Creativity	
Hypothesis 3	In a subjectively evaluated creativity contest, using peer evaluators, rather than manager evaluators, will decrease average idea usefulness.
Hypothesis 4	In a subjectively evaluated creativity contest, having participants compete for multiple small prizes, rather than one large prize, will decrease average idea creativity.
Research Question 1	In a subjectively evaluated creativity contest, do the job role of the contest's evaluator and the prize structure have a significant impact on the creativity of the best idea submissions?
Interactive Effects	
Research Question 2	In a subjectively evaluated creativity contest, does the job role of the contest's evaluator and the prize structure have a significant interaction effect on participation?
Research Question 3	In a subjectively evaluated creativity contest, does the job role of the contest's evaluator and the prize structure have a significant interaction effect on idea creativity?

3. Experimental Design

To test our hypotheses, we design a two-by-two experiment on Amazon's Mechanical Turk (M'turk) [Kittur, Chi, and Suh 2008, Paolacci, Chandler, and Ipeirotis 2010], which is an online labor market meant for crowd-sourcing workers to complete tasks like translations, research experiments, and changing pictures into machine readable data sources. The experiment takes the form of a contest in which participants are offered monetary incentives to submit their most creative ideas as to how researchers could improve the surveys conducted on Mechanical Turk. We incentivize participants to help solve a problem directly related to their everyday job on M'turk—developing a creative attention check question to improve the quality of completed surveys—just as a manager would provide incentives for solving job-specific problems in organizations [Harrison and List 2004, List 2011, Bloomfield et al. 2016].[5] This design choice allows participants to use their knowledge and experiences to develop and submit new creative ideas or to submit creative ideas that they had previously conceived during the course of their job. Adding in this level of realism makes this study a strong test of our hypotheses because we can measure real workplace contest participation behavior and the creativity of job-relevant ideas.

This advantageous aspect of our setting also differentiates our experiment from other experiments that (1) ask participants to complete creative tasks that are likely new to them, (2) do not give participants extended amounts of time to develop their ideas, or (3) do not allow researchers to capture organic participation rates [Kachelmeier et al. 2008, Gneezy et al. 2011, Ostermaier and Uhl 2020]. In addition, Mechanical Turk samples are nationally representative, making them a reasonable proxy for employees in the U.S., more so than student samples [Ross, Irani, Silberman, Zaldivar, and Tomlinson 2010, Berinsky, Huber, and Lenz 2012, Farrell, Grenier, and Leiby 2017]. The experiment is conducted in three sequential stages: participant screening, idea collection, and idea evaluation, which occur over the course of four months. Each stage is detailed below in the order that it occurs.

[5]This study is approved by the IRB at the University the study was conducted at (Study Number 2020-011) and is pre-registered with the American Economic Association (AEARCTR-0005443).

3.1. PARTICIPANT SCREENING

We develop a screening survey to identify online labor-market participants who had the necessary technological and tacit knowledge of M'turk to provide creative ideas about surveys conducted on the platform. We use Cloud Research, a third-party website not owned by Amazon, which allows us to first invite participants to complete the screening survey and then identify and invite only those who provided satisfactory responses to participate in the idea collection part of the experiment. We only invite participants with a task approval rate of 75% or above, who have completed over 1,000 human-intelligence tasks (hereafter tasks), and who are not on the Cloud Research Universal Exclude List, a list of IP addresses and Mechanical Turk IDs that have been marked as suspect in other tasks done on Cloud Research. The expected time to complete the screening survey is 10 minutes and participants are paid $1.33, which is comparable to other research surveys done on Mechanical Turk [Farrell et al. 2017, Bentley 2018, LaViers 2019].

Participants answer questions related to their demographic characteristics and their knowledge of and experience with Mechanical Turk. All potential participants see the same survey. In total 1,000 subjects completed the screening survey. Individuals are excluded from participating in the contest if they do not include their M'turk identification number in the appropriate response box on the initial survey, if they do not indicate that they wanted to be contacted for a future task, or if they do not answer every pre-screening question. Screening results in 941 participants, who are invited by email and via the Mechanical Turk platform to participate in the subsequent idea collection stage of our experiment. We provide more information about the individuals who are invited to participate in the contests in Section 4.

3.2. IDEA COLLECTION

In the idea collection stage of our experiment, we ask participants to submit to us their most creative ideas for attention checks that researchers could use to improve M'turk surveys. See Appendix A for the exact wording of the materials presented to the participants. An attention check is a question that tests survey takers' level of focus, often by quizzing them on what they just read or what

response they just provided or by asking them to do a quick, sometimes entertaining, engagement task to reduce the monotony of the survey. All of our participants report in the pre-survey that they are familiar with attention checks, which is not surprising since they have each completed at least 1,000 tasks. We randomly allocate each participant into one of four treatment cells (i.e., contests), in a two-by-two design, manipulating (1) the job role of the evaluator of the submitted ideas and (2) the number of prizes that participants compete for. To vary the role of the evaluator, half of the participants are told that the creativity of their ideas will be evaluated by a panel of fellow M'turkers, that is, peers with the same job role as them, and the other half are told that the creativity of their ideas will be evaluated by a panel of M'turk task requesters, that is, managers in charge of implementing tasks on the platform. Accordingly, we refer to these treatment cells as "peer evaluator" and "manager evaluator."[6] To vary the number of prizes, half of the participants (specifically, half in each of the two evaluator role treatment groups) are told that the person who submits the most creative idea will receive $100, and the other half are told that the contributors of the 10 most creative ideas will each receive $10. We refer to the former $\$100 \times 1$ prize structure as "single-prize" and to the latter as "multiple-prize."

Participants are allowed to submit multiple ideas, with a maximum of five per day, over the course of seven days. No participant hit the limit of 35 ideas, indicating that this was more than enough entries for each participant. Ideas need to be less than 750 characters, and each idea needs to be submitted as its own entry, so multiple ideas can not be submitted within a single submission. Participants are told that the creativity of each idea will be judged on two factors: "how novel (i.e., new) and how useful (i.e., realistic)" is the idea. After participants submitted each idea, they were asked two questions: "how long did you spend thinking about or working on this idea" and "how difficult or easy was it for you to think of this idea?" Responses to these questions allow us to estimate effort exertion. Contestants were invited to participate in the contests each day for a

[6]Potential conflicts of interest between peer evaluators and participants are important to consider. In our setting, no M'turk workers who are involved in evaluating the ideas are also participants in the contest, meaning there is no conflict of interest. In addition, all ideas are judged anonymously, so there can be no favoritism. In other settings, where participants and evaluators are peers at the same company and may be involved in multiple contests together, it may be prudent to explicitly work towards reducing any conflicts of interest and to take steps to make idea submission and evaluation processes as anonymous as possible.

week. Importantly, due to the multi-stage nature of our design, we can study the impact of contest design on *both* participation behavior and the creativity of the ideas submitted.

3.3. IDEA EVALUATION

After seven days of idea collection, the idea submission surveys are closed, and the idea evaluation begins. Peer evaluators are selected from Mechanical Turk, using the same screening survey with the same participant restrictions as was used for the experimental participants. The only change made to the screening survey was the addition of a free response question, "In 500 characters or less, explain why you would be a good judge for an M'turk based contest?" Fifty-two potential evaluators completed the screening survey, 46 of whom said they were interested in participating in another task. Of these, all three coauthors on this project independently rated the responses to the free response question on a scale of 1 to 10. Participants who earn an average score of 4 or above are invited to evaluate the ideas.[7] This results in 25 M'turk evaluators (i.e., peer evaluators).

Following other work on creativity, the evaluators judge the creativity of each idea on a scale from 0 to 100 points [Amabile 1983, Sethi, Smith, and Park 2001, Amabile 2012]. Evaluators also separately judge the novelty and usefulness of each idea on the same scale, which allows us to assess how the contest design choices impacted these features of creativity. To reduce the evaluators' decision fatigue, we randomly organize the ideas into 27 different surveys that evaluators can complete over time, with approximately 20 ideas per survey [Li and Sandino 2018]. M'turk evaluators are paid a flat fee of $2 for each survey they complete (i.e., each batch of approximately 20 ideas they evaluate). In addition, each survey has an attention check question at the end. Participants earn $1 if they correctly answer this question, resulting in a total pay of $3 per survey. Each survey takes approximately 15 minutes to complete. Participants who complete all 27 evaluation surveys make $81 total. To ensure high evaluation quality, participants are told that if they do not answer the attention check questions correctly, they will not be invited back to complete additional evalu-

[7] Evaluating the quality of the prospective evaluators' written responses allows us to screen for language skills and ensure that a person, not a bot, is responding. It also allows us to determine which participants are willing and able to put forth high effort while evaluating the ideas.

ations. Participants are sent three surveys on the first day. Their data is checked, their bonuses are paid, and then they are sent five more surveys the next day. This process repeats until participants complete all 27 surveys or are not invited back, due to either a lack of survey completion or poor evaluation quality. To mitigate order effects, the evaluators are sent the surveys in three different orders. Not every evaluator chooses to judge each idea, but all ideas are evaluated by at least 12 peer evaluators.

The manager evaluators are 7 M'turk task requesters who have experience designing research surveys and experiments on the platform. All of these task requesters have training in research methods from doctoral programs in North America and Europe. These manager evaluators are each paid $10 for completing the initial screening survey, and they have the same fixed plus variable pay ratio as the peer evaluators but at a higher rate. Their pay is set at $15 an hour, which results in a total payment of $145 if they complete all 27 surveys. They are sent a slightly modified screening survey, which does not ask for their experience as an M'turk participant. These evaluators are shown the same evaluation surveys as the peer evaluators, also in the same three patterns. Once again, not everyone chooses to judge every idea, but every idea is judged by at least six manager evaluators. This number of manager evaluators, while smaller than the number of peer evaluators, still exceeds or is similar in size to other creativity judging panels from prior literature [Kachelmeier et al. 2008, Amabile 2018, Li and Sandino 2018, Cardinaels, Dierynck, and Hu 2020]

3.4. SCORE AGGREGATION

An important consideration was how best to aggregate the various scores given to a single idea, considering the quantity and variety of evaluators. To ensure that evaluations by peers and those by managers are given similar weights in the aggregation, we began by identifying the evaluators in each group who evaluated the highest percentage of total ideas. Among the M'turk workers and task requesters who participated as evaluators, five in each group completed 26 or 27 of the 27 evaluation surveys. We label them "core evaluators." We then create an average creativity score

for each idea by taking the equal-weighted average of the creativity scores given to the idea by these core evaluators. Similarly, we create an average novelty (usefulness) score for each idea by taking the equal-weighted average of the novelty (usefulness) scores given to the idea by the core evaluators. We refer to this aggregation approach as *Core Evaluators*, and these are the scores we use throughout the paper.

To ensure our results are insensitive to the method chosen to aggregate scores, we consider several other methods of aggregation. The simplest is to take the equal-weighted average of all evaluations given to a particular idea. As at least 12 peer evaluators and six manager evaluators judged each idea, we have at least 18 evaluations for each idea. We call this approach, *All Evaluators*. One downside to this approach is that it overweights peer evaluations, as there are at least double the number of peer evaluators for each idea. We also aggregate scores using only evaluations from peer evaluators and using only evaluations from manager evaluators, and we refer to these approaches as *Peer Evaluators* and *Manager Evaluators*, respectively. Finally, to capture the possibility that contest participants may adjust the features of their ideas to appeal to their assigned evaluators, we also aggregate scores by equally weighting each evaluation from only the idea evaluators that we assigned to the participants, either peers or managers. We call this approach, *Assigned Evaluators*.[8]

4. Effect of Contest Design on Creative Output

Table 1 reports descriptive statistics about the submitted ideas and the participants. Panel A shows that, among the 527 ideas submitted, the average creativity score was 44.45, with the least creative idea receiving a creativity score of 4.00 and the most creative idea receiving a creativity score of 72.30. The average usefulness score was 49.41, with a low of 4.40 and a high of 76.20, and

[8]We show in Table B.1 in Appendix B that our main idea creativity results, which we discuss in Section 4.2, are robust to the use of each of these different measures. In Appendix C, we list the three ideas that were evaluated as the most creative in each of the four treatment cells, using the *Core Evaluators* scores. We also list the three ideas that were evaluated as the most useful and the three ideas that were evaluated as the most novel. To determine contest winners and allocate prize money, we used the *Assigned Evaluators* scores for creativity, so contest winners were determined based on their treatment cell designation, peer evaluators versus manager evaluators. Winning participants were paid immediately after all of the ideas were evaluated via the Mechanical Turk Bonus feature.

the average novelty score was 46.49, with a low of 3.70 and a high of 77.56. Though we told the evaluators that “creative ideas are those that are both useful and novel," we left it up to them to determine how much weight to give each of these features as they compiled a creativity score for each idea. Table 2 shows the correlations between the creativity scores and the novelty and usefulness scores of the ideas. The correlation between idea creativity and novelty, 0.959, is over twice as large as the correlation between idea creativity and usefulness, 0.432. This suggests that evaluators, in general, placed more weight on novelty as a determinant of creativity.

Panel A of Table 1 also shows that the average idea length was about 308 characters (including spaces). In addition, the average idea took 11.33 minutes to develop and share, and the average difficulty response value was 2.51, which is below the midpoint of the 1 to 5 response scale in which 1 denotes “extremely easy," 3 denotes “neither easy nor difficult," and 5 denotes “extremely difficult." We consider these proxies for effort exertion in greater detail in Section 5.

Panel B reports the demographic characteristics of the participants across each of the four treatment cells. This information was gathered during the screening survey. The average participant in our experiment is approximately 38 years old, and about 40% of the participants are women (59% are men and 1% are either non-binary or responded “prefer not to say"). The average participant has completed over 5,000 tasks, suggesting that the sample is composed of highly experienced M’turkers. The variation in demographics among our participants is analogous to what one would expect to find in a large, modern-day firm. Nearly all participants are high school graduates, and over 90% have completed some college classes, though they may not have attained a secondary degree yet. The far-right column reports p-values from tests of the hypothesis that the averages across the four cells are jointly equal. Across all demographic characteristics, we find no significant differences between conditions.

Panel C reports descriptive statistics of participation behavior. The first row reports the number of unique ideas that were submitted by participants in each of the four treatment cells. There were a total of 308 ideas submitted among participants in the peer evaluator treatment cells, and 219 ideas were submitted among participants in the manager evaluator treatment cells. Within the

single-prize treatment cells, 25 more ideas were submitted by participants with peer evaluators, and within the multiple-prize treatment cells, 64 more ideas were submitted by participants with peer evaluators. The second row reports the number of participants in each of the treatment cells who submitted at least one idea, and the bottom rows report the number of participants who submitted particular quantities of ideas. In total, 117 participants in peer evaluator treatment cells submitted at least one idea, 50 in the single-prize cell and 67 in the multiple-prize cell, and 111 participants in the manager evaluator treatment cells submitted at least one idea, 55 in the single-prize cell and 56 in the multiple-prize cell. Across all treatment cells, 713 participants chose to not contribute any ideas, 119 submitted exactly one idea, and 96 submitted between 2 and 5 ideas. Only 13 participants submitted six or more ideas, and the largest number of submissions from a single participant was 26.[9] In the following section, we formally document the impact of the contest design choices on participation behavior.

4.1. PARTICIPATION BEHAVIOR

In Section 2, we hypothesized that employees would participate more if they were told their ideas would be evaluated by their peers, rather than their managers. Similarly, we hypothesized that offering multiple prizes, as opposed to a single winner-take-all prize, would increase participation. To formally test these hypotheses, we use a sample of the 941 individuals who indicated on the screening survey that they wanted to complete an additional task, all of whom were invited to participate in one of the four contest treatments. We consider two measures of participation: a discrete measure of the number of ideas submitted by each participant (results presented in Column (1) of Table 3) and a binary measure of whether or not participants chose to enter the contest by submitting at least one idea (results presented in Column (2) of Table 3). In order to provide interpretable estimates of both the main effects and the interaction effect, simultaneously, we regress the participation measure on *Peer Evaluator*, *Multiple-Prize*, and their interaction, using effects coding. In all specifications we include the demographic characteristics of each participant—age, gender,

[9] As we discuss in the following section, our results are robust to various strategies of handling outliers.

years of education, and number of tasks completed on the platform—to control for the potential influence of these factors on participation.[10]

The results reported in Column (1) of Table 3 show that the coefficient on *Peer Evaluator* is positive and statistically significant at the 5% level when using our discrete measure of participation.[11] As for the binary measure of participation, the estimate on *Peer Evaluator* in Column (2) is statistically insignificant (p-value = 0.481). Our results indicate that peer evaluators elicit increased idea-sharing through a greater intensity of participation from contributing participants, but peer evaluators do not cause a significantly greater number of individuals to participate in the contest. The first result provides support for Hypothesis 1, whereas the second result is insignificant and does not provide support for the hypothesis.

When considering the idea submission counts in Panel C of Table 1, we see that removing the outlier, who submitted 26 ideas, from the sample reduces the number of ideas submitted in the *Peer Evaluator* $\times$ *Single-Prize* contest from 138 to 112, making it almost identical to the number of ideas submitted in the *Manager Evaluator* $\times$ *Single-Prize* contest. Examining the raw data suggests that in the absence of this outlier, the positive effect of peer evaluators on idea submission counts is only impactful when multiple prizes are offered. However, when we re-estimate our main regression specifications without the outlier, our inferences remain relatively unchanged. As shown in Column (1) of Panel A of Table B.2 in Appendix B, the positive estimate on *Peer Evaluator* remains significant if we remove the outlier from the sample (p-value = 0.088). The same holds true when we tag the outlier with a less extreme value of the number of ideas submitted (p-value = 0.054), as shown in Panel B.

As for the impact of prize structure on participation, the insignificant point estimates on *Multiple-Prize* in both Columns (1) and (2) of Table 3 suggest that participants, on average, did not participate more or less based on the prize structure of the contest. These results do not support Hy-

[10] As is expected in an experiment with random assignment to treatment cells, participants are balanced on these observable characteristics across the four treatment conditions, and our participation results are robust when we exclude these controls from the empirical tests.

[11] The results are similar if we exclude individuals who did not submit any ideas and then compare the number of submissions among contributors across the different treatment cells.

pothesis 2, but, as we discuss below, we find some evidence in support of this hypothesis when focusing on participants who are the most likely to feel they have a low win-likelihood. The insignificant estimates on *Multiple-Prize* $\times$ *Peer* in both Columns (1) and (2) indicate that there is not a significant interaction effect of evaluator role and prize structure on participation.

As discussed above, we find an insignificant main effect of *Multiple-Prize* on participation. This could be because the predicted effect does not exist at all or, alternatively, it could be driven by a lack of power from a weak experimental manipulation. To investigate further, we refine our participation analysis to a sample of the participant population where our theory predicts the effect of multiple-prize incentives would be the strongest: those who have a lower baseline belief in their win-likelihood. We proxy for this baseline belief by separating participants based on whether their demographic groups are under-represented or over-represented in creative and innovative endeavors. We argue that young males are over-represented in such pursuits, as young males are often cited by others as the most creative employees and entrepreneurs [Mason 2020]. This emphasis may be attributed to the fact that these workers are similar in demographics to famous innovators like Mark Zuckerburg, Evan Spiegel, and Jack Dorsey. Also consistent with this assertion, this demographic group is more likely to receive venture capital funding for their business ideas than are entrepreneurs in other demographic groups [Kornblum 2016, Lang and Van Lee 2020]. As another example, Hofstra et al. [2020] find that young males are more likely to have their PhD dissertations viewed as innovative, even when holding constant the actual creativity level of the papers. In order to separately examine these groups, we categorize participants as *Under-Represented* if they are either non-male or ≥ 50 years old, or both, and we categorize participants as *Over-Represented* if they are both male and < 50 years old.[12] We posit that participants in our contests who do not fit into the young-male demographic may have a lower baseline belief in their win-likelihood because they have less experience receiving positive recognition for their creative efforts. As such, we predict that multiple-prize incentives will have a positive effect on the participation behavior of these participants.

[12]Our results are similar if we instead use an age threshold of 40 years old.

In Table 4, we test for the predicted heterogeneous effect of multiple prizes on the participation of under-represented and over-represented demographic groups. Using our discrete measure, we re-estimate the model using only the *Under-Represented* sample (N = 497) in Column (1), while restricting our sample only to the *Over-Represented* sample (N = 444) in Column (2). The results tell a clear story: participants from under-represented demographics submit a greater number of ideas to multiple-prize contests, whereas participants from over-represented demographics do not. This is captured by the significant coefficient of 0.140 on *Multiple-Prize* in Column (1) and the insignificant coefficient of -0.090 in Column (2). These effects are significantly different, as indicated by the p-value of 0.036 in the penultimate row of the table. The significance of this difference in effect is also captured by the results in Column (3), wherein we interact the indicator variable, *Under-Represented*, into the model. The positive, significant coefficient on *Under-Represented* $\times$ *Multi.* implies that under-represented participants have a significantly more positive response to multiple-prize incentives than do over-represented participants.[13] In Columns (4)–(6), we repeat this analysis using the binary participation measure. However, in this case, the difference in the effect of *Multiple-Prize* on the participation of the two subsets of employees is not statistically significant (p-value = 0.403). Overall, we conclude that the use of peer evaluators can significantly increase the number of ideas submitted by all employees, and the use of multiple prizes can increase the number of ideas submitted by under-represented employees. However, we do not observe significant effects of prize structure for participation rates in the full sample.

As discussed above, we motivate our comparison of under-represented and over-represented subsets of employees with the notion that under-represented individuals are likely to have a lower baseline belief in their likelihood of winning subjectively evaluated creativity contests. To investigate the validity of this belief, we compare the responses of under- and over-represented participants to several questions about their own creativity. Participants were asked to rate how often they do the following activities: (1) Generate creative ideas; (2) Are innovative; (3) Develop adequate plans and schedules for the implementation of new ideas; (4) Investigate and secure funds needed

[13]The coefficients on *Peer Evaluator* in Columns (1) and (2) are similar in magnitude and are not statistically different from one another, as indicated by the p-value of 0.708 at the bottom of the table.

to implement new ideas; and (5) Promote and champion ideas to others. Participants could select from the following five frequencies: (1) Never; (2) Infrequently; (3) Neither often nor infrequently; (4) Often; and (5) Always.[14] We gathered these self-reported creativity ratings during the screening survey, in the first stage of our two-stage design. Figure 1 reports the differences in the average responses of under- versus over-represented participants. The first two creativity questions measure how creative and innovative participants believe themselves to be. There is no statistical difference in responses to these questions between under- and over-represented participants. This suggests that the reason under-represented participants submit more ideas in multiple-prize contests is not driven by them seeing themselves as less creative than others.

Although under-represented participants see themselves as just as creative as over-represented participants, they may believe themselves to be less likely to win the contest because of their lack of experience having their ideas formally recognized by others. To investigate this, we consider the three other self-reported creativity measures, specifically whether or not they have developed plans, received funding, or had experience promoting their own ideas in the past. We find that under-represented participants report significantly lower levels on the first two measures than do over-represented participants (at the 5% level). They also report that they are less likely to promote and champion their ideas to others (this difference is only marginally significant). Taken together, these differences in self-reported creativity suggest that under-represented participants have less experience having their ideas formally recognized and invested in by others. This lack of experience may cause them to feel they are less likely to win a winner-take-all contest, which drives their participation in multiple-prize contests.

4.2. IDEA CREATIVITY

Next, we test our hypotheses concerning the effect of contest design choices on the creativity of the submitted ideas, as well as how they impact the creative features of the ideas, their novelty and usefulness. To do this, we compile a dataset consisting of all 527 unique ideas submitted by

[14] These questions come from the innovative behavior measure scale developed by Scott and Bruce [1994].

the contest participants. We then perform ordinary least squares regressions where the dependent variable in Panel A of Table 5 is each idea's usefulness score, using the *Core Evaluators* aggregation approach, and the dependent variable in Panel B (Panel C) is each idea's novelty (creativity) score.[15] The specifications in Column (1) use all 527 ideas to identify average treatment effects, providing evidence as to how managers can impact the average creativity of the solicited ideas by manipulating a contest's evaluator role, prize structure, and both simultaneously. However, since the main focus of creativity contests is often eliciting a few ground-breaking ideas, managers are interested in knowing about contest design choices that positively impact the right-tail of the creativity distribution of ideas, with little concern as to what the affect is on average creativity across the entire distribution of idea submissions. To consider the impact of contest design choices on only highly creative ideas, we re-estimate our regression models using only the best ideas from each of the four treatment cells (i.e., the most creative (useful/novel) ideas when estimating the effect on creativity (usefulness/novelty)). The results of these sub-sample estimations are presented in Columns (2) and (3), which restrict the sample to the top 50 and 25 ideas from each treatment cell, respectively.[16]

4.2.1. Evaluator Role and Idea Creativity

We predict in Hypothesis 3 that using peers, rather than managers, to evaluate the ideas will lead to ideas that are less useful, on average. To test this hypothesis, we draw inference from the coefficient on *Peer Evaluator* in Column (1) of Panel A of Table 5. The negative, significant estimate suggests that participants submit ideas that are less useful, on average, when their peers, rather than their managers, evaluate the ideas. This result provides support for Hypothesis 3. The results in Columns (2) and (3), however, suggest that peer evaluators do not significantly impact the usefulness of the most useful ideas. So, while average idea usefulness declines when peer

[15] We control for participant demographics in all of the specifications, as we did in the participation analysis. This allows us to control for possible selection effects based on observable characteristics, as not all individuals chose to submit ideas to the contest.

[16] These sub-samples are similar in size to those in previous studies that defined the best ideas as those in the top quintile of ratings [Kachelmeier et al. 2008].

evaluators are used, the usefulness of the best ideas does not. These results indicate that managers should designate themselves as the evaluators of the ideas if they want to solicit ideas that are more useful on average (e.g., continuous improvement initiatives). If, however, they mainly care about the usefulness of the best ideas, then they should consider using peer evaluators, which lead to increased participation without reducing the usefulness of the best ideas.

While we have a formal hypothesis regarding the effect of evaluator role on idea usefulness, the impact on idea novelty is less obvious, making the effect on idea creativity more ambiguous. We posit that both peers and managers value novel ideas and that managers' extra appreciation for useful ideas does not necessarily mean participants will submit less novel ideas to managers (i.e., the creative features, novelty and usefulness, are not necessarily mutually exclusive).[17] In Panels B and C of Table 5, we show that, on average, evaluator role has no effect on idea novelty and creativity, respectively, as the point estimates on *Peer Evaluator* are small and statistically insignificant in Column (1). As discussed previously, and as shown in Table 2, creativity scores appear to be more highly correlated with novelty scores than with usefulness scores. So, even though peer evaluators lead to reduced usefulness, the null effect on novelty appears to drive the insignificant average treatment effect on creativity. Whereas evaluator role does not have a significant impact on the *average* novelty and creativity across the entire sample of ideas, we do find evidence that the novelty and creativity of the best submissions increase when peers, rather than managers, are assigned to evaluate the ideas. The positive, significant coefficients on *Peer Evaluator* in Column (2) of Panels B and C suggest that, among the top 50 idea submissions in each treatment cell, idea novelty and creativity increase significantly when peer evaluators are used. The effect on novelty persists when we consider only the top 25 ideas in each cell, suggesting that the most out-of-the-box ideas were submitted by participants in peer evaluator treatment cells. This finding is especially important for managers who are in search of ground-breaking ideas, as it suggests that the use of peer evaluators increases the novelty of the best ideas.

[17] In Appendix D, we formally analyze the differences in the weights that peer evaluators and manager evaluators give to usefulness and novelty when determining the creativity of an idea. We find evidence in support of our prediction that manager evaluators place significantly more weight on idea usefulness and less weight on idea novelty than do peer evaluators.

4.2.2. Number of Prizes and Idea Creativity

In Hypothesis 4, we predict that having participants compete for multiple prizes, rather than a single winner-take-all prize, will lead to decreased average creativity among the submitted ideas. To test this prediction, we estimate the effect of prize structure on the average creativity scores of the submitted ideas, as well as the effect on their usefulness and novelty scores. These effects are captured in the coefficients on *Multiple-Prize* in Column (1) of Table 5. In Panel A, we document a negative effect on usefulness, but the effect is not statistically significant. Similarly, the estimates in Columns (2) and (3) show that the effect of prize structure on idea usefulness is small and insignificant when considering only the most useful ideas. In Panels B and C, however, we document negative and statistically significant effects of offering multiple prizes on the average novelty and creativity of the ideas. These findings provide partial support for Hypothesis 4 and suggest that designing a subjectively evaluated contest with multiple winners of small prizes, rather than a single winner of a large prize, leads to reduced average idea novelty and creativity. In addition, we continue to find negative, statistically significant coefficients on *Multiple-Prize* when considering the novelty and creativity of the most novel and creative ideas in Columns (2) and (3). So, even among the best idea submissions, idea creativity decreases when participants compete for multiple smaller prizes, rather than a large winner-take-all prize. This is suggestive evidence that single-prize incentives, relative to multiple-prize incentives, prompt creative thinkers to exert additional effort in an attempt to submit the single most creative idea.

4.2.3. Interaction Effect and Idea Creativity

We find almost no evidence of a significant interaction effect of evaluator role and prize structure on idea creativity. In Panels A and B, we also find insignificant interaction effects on idea usefulness and novelty, respectively. These null effects are realized both when considering average usefulness and novelty among all ideas and the usefulness and novelty of only the best ideas. In Column (1) of Panel C, we find a marginally significant interaction effect on creativity among all ideas. This decrease may be due to peer evaluators causing reduced idea usefulness, while, simultaneously,

multiple-prize incentives causing reduced novelty. The interaction of the two might have made the contest feel especially informal and low-stakes to participants, leading to less creative ideas being submitted. But the interaction effect is insignificant in Columns (2) and (3), suggesting that an interaction effect of the two contest design choices on creativity does not exist among the most creative ideas.

5. Effort Exertion

We finish with an analysis that provides insight regarding the channels that link contest design choices to the creativity of the ideas submitted by participants. The idea creativity results suggest that, on average, the use of peer evaluators leads to reduced idea usefulness and the use of multiple-prize incentives leads to reduced idea novelty and creativity. These results are potentially due to reduced effort exertion of the participants, who might see the threshold necessary for a prize to be lower when peers evaluate the ideas or when multiple prizes are offered. To further test the notion that contest design affects effort exertion, we consider several additional proxies of participant effort. Each time a participant submitted an idea, they were asked two questions about that idea: (1) "Approximately how long did you spend thinking about or working on this idea?" and (2) "How difficult or easy was it for you to think of this idea?"[18] Participants responded to each of these questions immediately after submitting an idea, and these responses proxy for the effort exerted by each participant. As mentioned earlier, the average idea took over 11 minutes to develop and submit, and the average reported difficulty response was between "Somewhat easy" and "Neither easy nor difficult." As a third proxy for effort exertion, we consider the length of the submitted idea, as—all else equal—longer ideas require greater effort to submit.

To estimate the impact of contest design on effort exertion, we adopt similar specifications to those used in Table 5, but we set the dependent variable to be one of our three effort proxies: *Time*

[18] Answer options for the first question were as follows: "I did not spend any time at all," "Less than 5 minutes," "More than 5 minutes, less than 15 minutes," "More than 15 minutes, less than 30 minutes," "More than 30 minutes, less than 1 hour," and "More than 1 hour." Answer options for the second question were as follows: "Extremely easy," "Somewhat easy," "Neither easy nor difficult," "Somewhat difficult," and "Extremely difficult."

Spent Thinking in Panel A, *Difficulty of Thinking* in Panel B, and *Idea Length* in Panel C. Table 6 displays the results. In Column (1), we regress the effort proxies on *Peer Evaluator*, *Multiple-Prize*, their interaction, and on the participant's demographic controls. Across all three effort proxies, we estimate a negative and statistically significant relation between *Multiple-Prize* and effort exertion. Specifically, when participants compete for several smaller prizes, as opposed to one large prize, they spend less time thinking of the idea, feel the idea is easier to conceive, and submit ideas that are significantly shorter. The negative effect of multiple-prize incentives on effort exertion continues to exist when we consider only the best ideas, as indicated in Columns (2) and (3). The relation between evaluator role and effort exertion is more nuanced, as we estimate negative effects on *Time Spent Thinking* and *Idea Length*, but positive effects on *Difficulty of Thinking*. In addition, the estimates on *Peer Evaluator* do not consistently extend into the subsets that only consider the most creative ideas.

These results provide meaningful insights for managers. If a manager's goal is to increase employee participation or push for continuous improvement, where every idea matters, then designing a contest with peer evaluators or multiple-prize incentives may be the best approach. It appears, however, that this design choice might lead to reduced average effort exertion among participants. As such, if the manager wants participants to exert increased effort to devise highly creative ideas, then higher-powered incentives are likely warranted.

6. Conclusion

We conduct an experiment to examine employees' willingness to develop and share their value-enhancing ideas when motivated by a contest that rewards creativity. We hypothesize and find that participation decisions and the creativity of the ideas submitted by participants can be substantially influenced by changes in the job role of the contest's evaluators and in the number of prizes that participants compete for. We find that employees submit more ideas when their peers, rather than their managers, evaluate their ideas, though the number of individuals who contribute ideas is unaffected. When studying the creative features of the submitted ideas, we find that employees submit

ideas that are significantly more useful, on average, when the ideas are evaluated by a manager, whereas the novelty of the best ideas increases when the ideas are evaluated by a peer. This leads to increased creativity among the best ideas when peer evaluators are used. In addition, we show that the creativity of the submitted ideas, including the best ideas, decreases when multiple-prize incentives are used, rather than a large winner-take-all price. This reduction in creativity is again driven by reduced novelty.

Our setting has limitations that provide opportunities for future research. In discussions about creativity, there is often a distinction made between the “light bulb moment," when an employee first conceives a creative idea, and the further development and sharing of the idea. In our setting, it very well may be true that the contest incentives enhanced the inception of creative ideas, but our experimental design does not allow us to precisely delineate between ideas that originated before or after the onset of the contest. We leave it to future research to examine the impact of creativity contest design on the initial stage of creative idea development. Furthermore, our experimental design varies the level of similarity between idea contributors and evaluators based on whether the two have the same job role within the organization. Subsequent studies could consider how similarities and dissimilarities between the contributors and evaluators along other dimensions, such as demographic characteristics, affect the quantity and features of creative output. Lastly, participants in our setting submitted ideas anonymously, so public recognition was not likely motivating their participation and effort exertion decisions. In other settings, reputational considerations may be substantial and the desire to receive recognition might influence how contest participants respond to manager versus peer evaluators.

References

ADAMCZYK, S.; A. C. BULLINGER; AND K. M. MÖSLEIN. "Innovation contests: A review, classification and outlook." *Creativity and Innovation Management* 21 (2012): 335–360.

AMABILE, T. M. "Social psychology of creativity: A consensual assessment technique." *Journal of personality and social psychology* 43 (1982): 997.

AMABILE, T. M. "The social psychology of creativity: A componential conceptualization." *Journal of personality and social psychology* 45 (1983): 357.

AMABILE, T. M. "Componential theory of creativity." *Harvard Business School* 12 (2012): 1–10.

AMABILE, T. M. Creativity in context: Update to the social psychology of creativity. Routledge (2018).

BATSON, C. D.; S. EARLY; AND G. SALVARANI. "Perspective taking: Imagining how another feels versus imaging how you would feel." *Personality and social psychology bulletin* 23 (1997): 751–758.

BENTLEY, J. W. "Challenges with Amazon Mechanical Turk research in accounting." *Available at SSRN 2924876* (2018).

BERINSKY, A. J.; G. A. HUBER; AND G. S. LENZ. "Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk." *Political analysis* 20 (2012): 351–368.

BLOOMFIELD, R.; M. W. NELSON; AND E. SOLTES. "Gathering data for archival, field, survey, and experimental accounting research." *Journal of Accounting Research* 54 (2016): 341–395.

BOUDREAU, K. J.; N. LACETERA; AND K. R. LAKHANI. "Incentives and problem uncertainty in innovation contests: An empirical analysis." *Management science* 57 (2011): 843–863.

BRADLER, C.; S. NECKERMANN; AND A. J. WARNKE. "Incentivizing Creativity: A Large-Scale Experiment with Performance Bonuses and Gifts." *Journal of Labor Economics* 37 (2019): 793–851.

CARDINAELS, E.; B. DIERYNCK; AND W. HU. "Rejection, Reward Type, and Employee Creativity: When Chocolate Is Better than Cash." *Georgia Tech Scheller College of Business Research Paper* (2020).

CHANG, A. "Deep Inside a Facebook Hackathon, Where the Future of Social Media Begins." *Wired* (2012).

COHEN, C.; T. R. KAPLAN; AND A. SELA. "Optimal rewards in contests." *The RAND Journal of Economics* 39 (2008): 434–451.

CONNELLY, B. L.; L. TIHANYI; T. R. CROOK; AND K. A. GANGLOFF. "Tournament theory: Thirty years of contests and competitions." *Journal of Management* 40 (2014): 16–47.

DEAN, C. AND E. IMAGE. "If you have a problem, ask everyone." *New York Times* 22 (2008).

DECHENAUX, E.; D. KOVENOCK; AND R. M. SHEREMETA. "A survey of experimental research on contests, all-pay auctions and tournaments." *Experimental Economics* 18 (2015): 609–669.

EPLEY, N.; B. KEYSAR; L. VAN BOVEN; AND T. GILOVICH. "Perspective taking as egocentric anchoring and adjustment." *Journal of personality and social psychology* 87 (2004): 327.

FAIRBANK, J.; W. SPANGLER; AND S. D. WILLIAMS. "Motivating creativity through a computer-mediated employee suggestion management system." *Behaviour & Information Technology* 22 (2003): 305–314.

FAIRBANK, J. F. AND S. D. WILLIAMS. "Motivating creativity and enhancing innovation through employee suggestion system technology." *Creativity and innovation management* 10 (2001): 68–74.

FARRELL, A. M.; J. H. GRENIER; AND J. LEIBY. "Scoundrels or stars? Theory and evidence on the quality of workers in online labor markets." *The Accounting Review* 92 (2017): 93–114.

GARICANO, L. AND L. RAYO. "Relational knowledge transfers." *American Economic Review* 107 (2017): 2695–2730.

GNEEZY, U.; S. MEIER; AND P. REY-BIEL. "When and why incentives (don't) work to modify behavior." *Journal of Economic Perspectives* 25 (2011): 191–210.

GRABNER, I. "Incentive system design in creativity-dependent firms." *The Accounting Review* 89 (2014): 1729–1750.

GRANT, A. M. AND J. W. BERRY. "The necessity of others is the mother of invention: Intrinsic and prosocial motivations, perspective taking, and creativity." *Academy of management journal* 54 (2011): 73–96.

GREEN, J. R. AND N. L. STOKEY. "A comparison of tournaments and contracts." *Journal of Political Economy* 91 (1983): 349–364.

HARRISON, G. W. AND J. A. LIST. "Field experiments." *Journal of Economic literature* 42 (2004): 1009–1055.

HOFSTRA, B.; V. V. KULKARNI; S. M.-N. GALVEZ; B. HE; D. JURAFSKY; AND D. A. MCFARLAND. "The diversity–innovation paradox in science." *Proceedings of the National Academy of Sciences* 117 (2020): 9284–9291.

JEPPESEN, L. B. AND K. R. LAKHANI. "Marginality and problem-solving effectiveness in broadcast search." *Organization science* 21 (2010): 1016–1033.

KACHELMEIER, S. J.; B. E. REICHERT; AND M. G. WILLIAMSON. "Measuring and motivating quantity, creativity, or both." *Journal of Accounting Research* 46 (2008): 341–373.

KACHELMEIER, S. J.; L. W. WANG; AND M. G. WILLIAMSON. "Incentivizing the creative process: From initial quantity to eventual creativity." *The Accounting Review* 94 (2019): 249–266.

KACHELMEIER, S. J. AND M. G. WILLIAMSON. "Attracting creativity: The initial and aggregate effects of contract selection on creativity-weighted productivity." *The Accounting Review* 85 (2010): 1669–1691.

KIERSZ, A. "An eminent mathematician claims to have solved one of math's greatest mysteries — and it's one of 6 problems with a $1 million prize." *Business Insider* (2008).

KITTUR, A.; E. H. CHI; AND B. SUH. "Crowdsourcing user studies with Mechanical Turk." In "Proceedings of the SIGCHI conference on human factors in computing systems," (2008): pages 453–456.

KORNBLUM, R. "Older founders don't need VCs." *Business Insider* (2016).

LANG, I. AND R. VAN LEE. "Institutional Investors Must Help Close the Race and Gender Gaps in Venture Capital." *Harvard Business Review* (2020).

LAVIERS, L. "The Effect of Pay Transparency on Narcissists: Can Personality Type Predict Reciprocity?" *Available at SSRN 3487157* (2019).

LAZEAR, E. P. AND S. ROSEN. "Rank-order tournaments as optimum labor contracts." *Journal of political Economy* 89 (1981): 841–864.

LI, S. X. AND T. SANDINO. "Effects of an information sharing system on employee creativity, engagement, and performance." *Journal of Accounting Research* 56 (2018): 713–747.

LIST, J. A. "Why economists should conduct field experiments and 14 tips for pulling one off." *Journal of Economic perspectives* 25 (2011): 3–16.

MASON, H. "Should VCs invest in young or old founders? Using hard data, we find out." *Medium* (2020).

MEAD, G. H. Mind, self and society, volume 111. Chicago University of Chicago Press. (1934).

MENZEL, A. "Knowledge exchange and productivity spill-overs in Bangladeshi garment factories." *CERGE-EI Working Paper Series* (2017).

MOLDOVANU, B. AND A. SELA. "The optimal allocation of prizes in contests." *American Economic Review* 91 (2001): 542–558.

MORRISON, A. D. AND W. J. WILHELM JR. "Partnership firms, reputation, and human capital." *American Economic Review* 94 (2004): 1682–1692.

OSTERMAIER, A. AND M. UHL. "Performance evaluation and creativity: Balancing originality and usefulness." *Journal of Behavioral and Experimental Economics* (2020): 101552.

PAOLACCI, G.; J. CHANDLER; AND P. G. IPEIROTIS. "Running experiments on amazon mechanical turk." *Judgment and Decision making* 5 (2010): 411–419.

PARKER, S. K. AND C. M. AXTELL. "Seeing another viewpoint: Antecedents and outcomes of employee perspective taking." *Academy of Management Journal* 44 (2001): 1085–1100.

REID, S. E. AND U. DE BRENTANI. "The fuzzy front end of new product development for discontinuous innovations: A theoretical model." *Journal of product innovation management* 21 (2004): 170–184.

ROSS, J.; L. IRANI; M. S. SILBERMAN; A. ZALDIVAR; AND B. TOMLINSON. "Who are the crowdworkers? Shifting demographics in Mechanical Turk." In "CHI'10 extended abstracts on Human factors in computing systems," pages 2863–2872 (2010): .

RUNCO, M. A. AND G. J. JAEGER. "The standard definition of creativity." *Creativity research journal* 24 (2012): 92–96.

SANDVIK, J. J.; R. E. SAOUMA; N. T. SEEGERT; AND C. T. STANTON. "Workplace knowledge flows." *The Quarterly Journal of Economics* (2020).

SCOTT, S. G. AND R. A. BRUCE. "Determinants of innovative behavior: A path model of individual innovation in the workplace." *Academy of management journal* 37 (1994): 580–607.

SETHI, R.; D. C. SMITH; AND C. W. PARK. "Cross-functional product development teams, creativity, and the innovativeness of new consumer products." *Journal of marketing research* 38 (2001): 73–85.

SIEGEL, R. "All-pay contests." *Econometrica* 77 (2009): 71–92.

TERWIESCH, C. AND Y. XU. "Innovation contests, open innovation, and multiagent problem solving." *Management science* 54 (2008): 1529–1543.

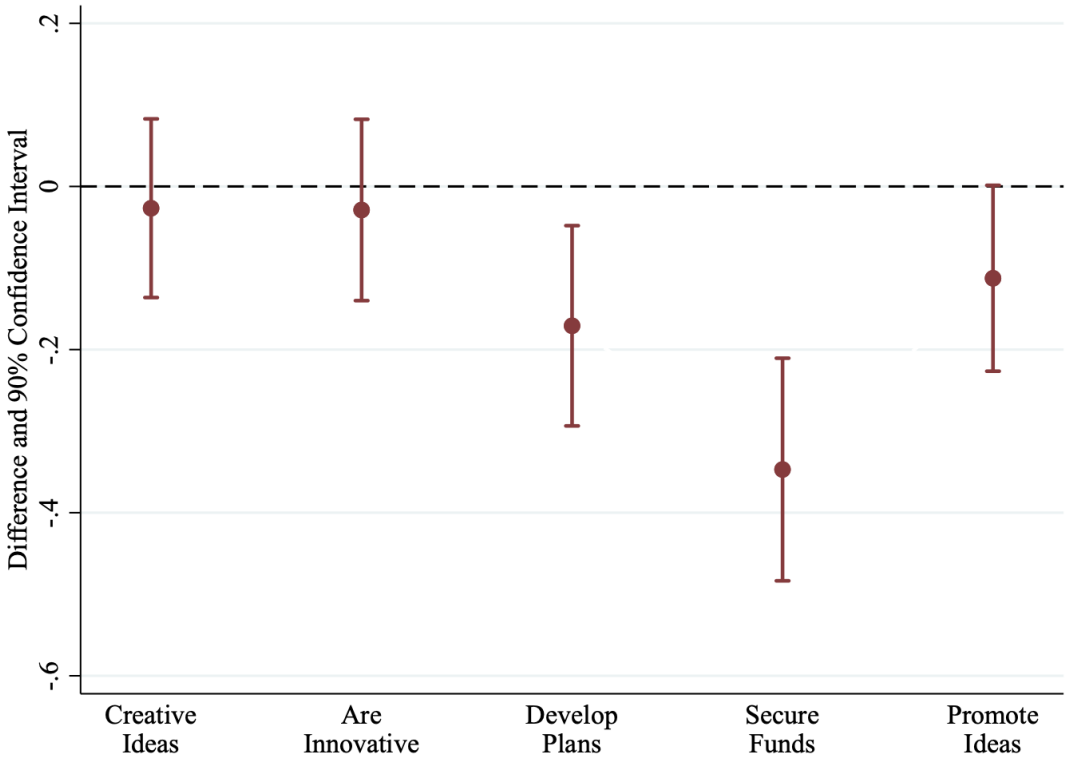
UNSWORTH, K. "Unpacking creativity." *Academy of management review* 26 (2001): 289–297.

VAN DIJK, C. AND J. VAN DEN ENDE. "Suggestion systems: transferring employee creativity into practicable ideas." *R&D Management* 32 (2002): 387–395.

WALKER, H. "Internal Company Hackathons - 5 successful examples." *Gravitywell* (2019).

Figures and Tables

Figure 1: *Differences in Self-Reported Creativity Between Sub-Samples*



Notes. This figure plots the differences in self-reported creativity measures between participants who are under-represented in creative endeavors and those who are over-represented. So, a negative average difference indicates that under-represented participants report a lower average creativity measure than do over-represented participants. Participants were asked to rate how often they do the following activities: (1) Generate creative ideas; (2) Are innovative; (3) Develop adequate plans and schedules for the implementation of new ideas; (4) Investigate and secure funds needed to implement new ideas; and (5) Promote and champion ideas to others. Participants could select from the following five frequencies: (1) Never; (2) Infrequently; (3) Neither often nor infrequently; (4) Often; and (5) Always. These questions come from the innovative behavior measure scale developed by Scott and Bruce [1994]. We also plot 90% confidence intervals around each difference.

Table 1: *Descriptive Statistics for the Main Variables of Interest*

Panel A: Descriptive Statistics of the Ideas Shared

	N	Mean	SD	Minimum	Maximum
Creativity	527	44.45	13.54	4.00	72.30
Usefulness	527	49.41	12.67	4.40	76.20
Novelty	527	46.49	14.29	3.70	77.56
Idea Length (characters)	527	307.94	232.84	23	1,629
Time Spent Thinking (min.)	527	11.33	13.74	0.00	60.00
Difficulty of Thinking	527	2.51	1.04	1.00	5.00

Panel B: Descriptive Statistics of Participant Characteristics

	Peer Evaluators		Manager Evaluators		
	Single-Prize ($100 x 1)	Multiple-Prize ($10 x 10)	Single-Prize ($100 x 1)	Multiple-Prize ($10 x 10)	Balance (*p*-value)
Age	38.09	38.34	39.06	38.35	0.838
Female	0.42	0.36	0.44	0.40	0.311
Years of Education	15.17	15.23	15.45	15.19	0.404
Number of HITs (00s)	50.71	54.62	50.80	55.60	0.442
Number of Participants	235	235	235	236	

Panel C: Descriptive Statistics of Participation Behavior

	Peer Evaluators		Manager Evaluators		
	Single-Prize ($100 x 1)	Multiple-Prize ($10 x 10)	Single-Prize ($100 x 1)	Multiple-Prize ($10 x 10)	Total
Ideas Submitted	138	170	113	106	527
Participants with $\geq$1 Idea	50	67	55	56	228
Participants with 0 Ideas	185	168	180	180	713
Participants with 1 Ideas	24	36	32	27	119
Participants with 2–5 Ideas	22	26	20	28	96
Participants with 6–10 Ideas	3	3	2	1	9
Participants with 11–15 Ideas	0	2	1	0	3
Participants with 16+ Ideas	1	0	0	0	1

Notes. Panel A is based on the sample of the 527 unique ideas submitted across all four contests. Panel B is based on the sample of the 941 individuals who were invited to participate in the contest after completing the screening survey and indicating their interest in being invited back to complete another task. Panel C reports the number of ideas submitted in each treatment cell and the number of participants who submitted particular quantities of ideas.

Table 2: *Correlation Table*
(two-tailed p-values in parentheses)

	Creativity	Usefulness	Novelty	Length	Time	Difficulty
Creativity	1.000					
Usefulness	0.432	1.000				
	(0.001)					
Novelty	0.959	0.282	1.000			
	(0.001)	(0.001)				
Idea Length	0.224	0.124	0.187	1.000		
	(0.001)	(0.005)	(0.001)			
Time Spent Thinking	-0.006	0.033	-0.043	0.328	1.000	
	(0.890)	(0.450)	(0.327)	(0.001)		
Difficulty of Thinking	0.149	0.076	0.127	0.091	0.362	1.000
	(0.001)	(0.081)	(0.004)	(0.037)	(0.001)	

Notes. Correlations across the sample of 527 unique ideas. *p*-values are denoted in parentheses to capture the statistical significance of the correlations. Values of 0.001 are used for all values less than or equal to 0.001.

Table 3: *Effect of Design Choices on Participation Behavior*
(two-tailed p-values in parentheses)

Dependent Variable:	Number of Ideas Submitted	Submitted at Least One Idea
	(1)	(2)
Peer Evaluator	0.103**	0.010
	(0.048)	(0.481)
Multiple-Prize	0.031	0.019
	(0.573)	(0.161)
Multiple-Prize × Peer	0.037	0.015
	(0.471)	(0.274)
Demographic Controls	✓	✓
Adj. R-Square	0.003	0.035
Observations	941	941

Notes. The dependent variable in Column (1) is equal to the number of ideas submitted by each participant, which captures participation. The value of this dependent variable is set to zero for individuals who were invited to compete in the contest but who chose not to submit an idea. The dependent variable in Column (2) is equal to one for participants who submitted at least one idea, and zero for individuals who were invited to compete in the contest but who chose not to submit an idea. In both specifications, we regress the dependent variable on treatment cell indicator variables, *Multiple-Prize* and *Peer Evaluator*, and on the interaction term between the two. We use effect coding, as opposed to dummy coding, to provide interpretable estimates of both the main effects and the interaction effects. We include demographic control variables for each participant's age, gender, education, and number of tasks completed in all specifications. Regression specifications include a constant, but we do not report the constant for brevity. We estimate heteroscedasticity-consistent standard errors and report two-tailed p-values in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 4: *Effects of Design Choices on Participation Behavior*
(two-tailed p-values in parentheses)

	Number of Ideas Submitted			Submitted at Least One Idea		
	Under-Represented	Over-Represented	All Workers	Under-Represented	Over-Represented	All Workers
	(1)	(2)	(3)	(4)	(5)	(6)
Peer Evaluator	0.124*	0.083	0.083	0.018	-0.001	-0.001
	(0.068)	(0.343)	(0.343)	(0.385)	(0.954)	(0.951)
Multiple-Prize	0.140**	-0.090	-0.088	0.030	0.007	0.007
	(0.039)	(0.304)	(0.317)	(0.137)	(0.693)	(0.689)
Multiple-Prize × Peer	0.069	0.009	0.007	0.008	0.025	0.025
	(0.296)	(0.914)	(0.938)	(0.681)	(0.185)	(0.189)
Under-Represented × Peer			0.036			0.018
			(0.746)			(0.520)
Under-Represented × Multi.			0.224**			0.022
			(0.041)			(0.432)
Under-Represented × Multi. × Peer			0.066			-0.015
			(0.551)			(0.573)
Under-Represented			0.050			0.041
			(0.672)			(0.180)
Demographic Controls	✓	✓	✓	✓	✓	✓
Adj. R-Square	0.021	-0.008	0.006	0.026	0.013	0.032
Observations	497	444	941	497	444	941
H_0: Peer Evaluator	0.708			0.498		
H_0: Multiple-Prize	0.036			0.403		
H_0: Multiple-Prize × Peer	0.580			0.542		

Notes. The dependent variable in Columns (1)–(3) is equal to the number of ideas submitted by each participant, which captures participation. The value of this dependent variable is set to zero for individuals who were invited to compete in the contest but who chose not to submit an idea. The dependent variable in Columns (4)–(6) is equal to one for participants who submitted at least one idea, and zero for individuals who were invited to compete in the contest but who chose not to submit an idea. In Columns (1) and (4), we restrict the sample to participants who are under-represented in creative endeavors—those who are either non-male and/or $\geq$ 50 years old. In Columns (2) and (5), we restrict the sample to participants who are over-represented in creative endeavors—those who are male and $<$ 50 years old. In both specifications, we include the interaction term between *Multiple-Prize* and *Peer Evaluator*, and we use effect coding, as opposed to dummy coding, to provide interpretable estimates of both the main effects and the interaction effects. We also include demographic control variables for each participant's age, education, and number of tasks completed. In the specifications in Columns (3) and (6), we use the full sample of participants and interact the dummy variable *Under-Represented*—which equals one if the participant is either non-male and/or $\geq$ 50 years old, and zero otherwise—into the model. All specifications include a constant, but we do not report the constant for brevity. We estimate heteroscedasticity-consistent standard errors and report two-tailed *p*-values in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively. The *p*-values in the bottom three rows report the significance of the difference between the labeled coefficient in the Columns (1) and (2) and Columns (4) and (5).

Table 5: *Effect of Design Choices on Idea Creativity*
(two-tailed p-values in parentheses)

Panel A: Usefulness of the Idea

	All Ideas	Top 50	Top 25
	(1)	(2)	(3)
Peer Evaluator	-1.401***	0.052	0.134
	(0.010)	(0.860)	(0.669)
Multiple-Prize	-0.428	0.160	-0.028
	(0.432)	(0.573)	(0.928)
Multiple-Prize × Peer	-0.743	0.090	-0.223
	(0.161)	(0.762)	(0.467)
Demographic Controls	✓	✓	✓
Adj. R-Square	0.058	-0.024	-0.051
Observations	527	200	100

Panel B: Novelty of the Idea

	All Ideas	Top 50	Top 25
	(1)	(2)	(3)
Peer Evaluator	-0.187	1.624***	1.114**
	(0.759)	(0.000)	(0.015)
Multiple-Prize	-1.814***	-1.976***	-2.807***
	(0.003)	(0.000)	(0.000)
Multiple-Prize × Peer	-1.000	0.476	0.267
	(0.103)	(0.265)	(0.578)
Demographic Controls	✓	✓	✓
Adj. R-Square	0.051	0.174	0.317
Observations	527	200	100

Panel C: Creativity of the Idea

	All Ideas	Top 50	Top 25
	(1)	(2)	(3)
Peer Evaluator	-0.510	1.046***	0.018
	(0.376)	(0.008)	(0.966)
Multiple-Prize	-1.567***	-1.951***	-1.908***
	(0.007)	(0.000)	(0.000)
Multiple-Prize × Peer	-0.972*	0.397	0.513
	(0.095)	(0.322)	(0.224)
Demographic Controls	✓	✓	✓
Adj. R-Square	0.053	0.128	0.219
Observations	527	200	100

Notes. In Column (1), we use a dataset consisting of all 527 unique ideas shared by the contest participants. We then perform ordinary least squares regressions to estimate the effects of contest design on idea creativity. We regress the dependent variable on treatment cell indicator variables, *Multiple-Prize* and *Peer Evaluator*, and on the interaction term between the two. We use effect coding, as opposed to dummy coding, to provide interpretable estimates of both the main effects and the interaction effects. The dependent variable in Panel A, B, and C is an idea's average usefulness, novelty, and creativity score, respectively, using the *Core Evaluator* aggregation. In Columns (2) and (3), we restrict the sample to the highest ranked 50 and 25 ideas, respectively, in each of the four treatment cells based on the scoring metric of focus. We control for each participant's age, gender, education, and number of tasks completed in all specifications. All regression specifications include a constant, but we do not report the constant for brevity. We estimate heteroscedasticity-consistent standard errors and report two-tailed *p*-values in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively

Table 6: *Effect of Design Choices on Effort Exertion*
(two-tailed p-values in parentheses)

Panel A: Time Spent Thinking

	All Ideas	Top 50	Top 25
	(1)	(2)	(3)
Peer Evaluator	-1.264**	-1.312	-0.014
	(0.050)	(0.152)	(0.993)
Multiple-Prize	-1.894***	-2.459**	-3.353*
	(0.003)	(0.029)	(0.050)
Multiple-Prize × Peer	0.682	0.568	0.618
	(0.287)	(0.601)	(0.697)
Demographic Controls	✓	✓	✓
Adj. R-Square	0.036	0.071	0.061
Observations	527	200	100

Panel B: Difficulty of Thinking

	All Ideas	Top 50	Top 25
	(1)	(2)	(3)
Peer Evaluator	0.119**	0.202***	0.276**
	(0.011)	(0.008)	(0.016)
Multiple-Prize	-0.217***	-0.253***	-0.265**
	(0.000)	(0.004)	(0.040)
Multiple-Prize × Peer	0.026	-0.011	-0.052
	(0.578)	(0.885)	(0.600)
Demographic Controls	✓	✓	✓
Adj. R-Square	0.048	0.051	0.117
Observations	527	200	100

Panel C: Idea Length

	All Ideas	Top 50	Top 25
	(1)	(2)	(3)
Peer Evaluator	-37.404***	-20.325	-7.255
	(0.000)	(0.225)	(0.778)
Multiple-Prize	-31.469***	-48.968***	-20.463
	(0.002)	(0.006)	(0.440)
Multiple-Prize × Peer	-10.782	6.297	56.313**
	(0.298)	(0.717)	(0.028)
Demographic Controls	✓	✓	✓
Adj. R-Square	0.055	0.043	0.138
Observations	527	200	100

Notes. The specifications in this table mirror those in Table 5, but with different dependent variables that proxy for effort exertion. In Panel A, we regress *Time Spent Thinking* on indicator variables for *Multiple-Prize* and *Peer Evaluator*, and on their interaction. We control for each participant's age, gender, education, and number of tasks completed in all specifications. In Panel B we consider *Difficulty of Thinking*, and in Panel C we consider *Idea Length*. Regression specifications include a constant, but we do not report the constant for brevity. We estimate heteroscedasticity-consistent standard errors and report two-tailed p-values in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively

A. Documentation

A.1. INVITATION EMAIL:

Welcome to The Creativity Contest! You've been selected to be a part of a contest. We've selected you based on the knowledge and experience you demonstrated about Mechanical Turk in our screening survey. The goal of the contest is to find the most creative ideas to help us improve our surveys. The ten most creative ideas will each receive a $10 bonus. [or] The most creative idea will receive a $100 bonus. The creativity of each idea will be judged by a panel of M'turk workers. [or] by a panel of M'turk HIT requesters. To find out more open the HIT! Here is the link: http://app.cloudresearch.com/Router/Mturk/233415 Thank you The BLS Research Project Team

A.2. INSTRUCTIONS:

Introduction: We are looking for your creative ideas to help us improve our M'turk based surveys.

One major problem we have is that some participants stop paying attention and don't give us high quality answers throughout the entire survey. We are seeking your help to fix this problem. We want to know what you think the most creative way to get people to pay attention is. In order to gather the ideas, we are running a contest and inviting you to be a part of it.

Contest Rules: The creativity of each idea will be judged on two factors: how novel (i.e., new) and how useful (i.e., realistic) is the idea? Each idea submitted will be judged anonymously by a panel of M'turk workers. [or] by a panel of M'turk HIT requesters. The 10 most creative ideas will each receive a $10 bonus. [or] The most creative idea will receive a $100 bonus. Take your time to think of your ideas, this survey is open to you until the 23rd of July. It won't fill up or close. If you have more than one idea, you can submit multiple ideas in one survey or you can submit this survey multiple times. Each idea you submit increases your chances of being a winner! Approximately 240 other M'turkers have been invited to participate in this contest. To submit your ideas and enter the contest, click next!

A.3. IDEA ENTRY:

Researchers want M'turk participants to pay more attention during surveys so that they provide more high quality answers.

Using 750 characters or less, What is a creative idea that you have for an attention check on M'turk?

Please only enter one idea in this text box. If you have more than one idea, you can enter it on the next page.

A.4. POST QUESTIONS:

Approximately how long did you spend thinking about or working on this idea?

- I did not spend any time at all.
- Less than 5 minutes.

- More than 5 minutes, less than 15 minutes.
- More than 15 minutes, less than 30 minutes.
- More than 30 minutes, less than 1 hour.
- More than 1 hour.

How difficult or easy was it for you to think of this idea?

- Extremely difficult
- Somewhat difficult
- Neither easy nor difficult
- Somewhat easy
- Extremely easy

B. Robustness

Table B.1: *Robustness of Effects on Idea Creativity Across Score Aggregation Approaches*
(two-tailed p-values in parentheses)

Panel A: Usefulness of the Idea

	(1)	(2)	(3)	(4)	(5)
Peer Evaluator	-1.401***	-1.561***	-1.511***	-1.587***	-1.561***
	(0.010)	(0.002)	(0.006)	(0.004)	(0.002)
Multiple-Prize	-0.428	-0.279	-0.007	-0.378	-0.279
	(0.432)	(0.584)	(0.991)	(0.495)	(0.584)
Multiple-Prize × Peer	-0.743	-1.417***	-0.544	-1.704***	-1.417***
	(0.161)	(0.004)	(0.321)	(0.002)	(0.004)
Core Evaluators	✓				
All Evaluators		✓			
Manager Evaluators			✓		
Peer Evaluators				✓	
Assigned Evaluators					✓
Demographic Controls	✓	✓	✓	✓	✓
Adj. R-Square	0.058	0.076	0.046	0.076	0.076
Observations	527	527	527	527	527

Panel B: Novelty of the Idea

	(1)	(2)	(3)	(4)	(5)
Peer Evaluator	-0.187	-0.547	-0.500	-0.566	-0.547
	(0.759)	(0.285)	(0.429)	(0.276)	(0.285)
Multiple-Prize	-1.814***	-1.588***	-1.664***	-1.583***	-1.588***
	(0.003)	(0.002)	(0.009)	(0.003)	(0.002)
Multiple-Prize × Peer	-1.000	-1.626***	-0.642	-1.947***	-1.626***
	(0.103)	(0.002)	(0.311)	(0.000)	(0.002)
Core Evaluators	✓				
All Evaluators		✓			
Manager Evaluators			✓		
Peer Evaluators				✓	
Assigned Evaluators					✓
Demographic Controls	✓	✓	✓	✓	✓
Adj. R-Square	0.051	0.072	0.054	0.078	0.072
Observations	527	527	527	527	527

Panel C: Creativity of the Idea

	(1)	(2)	(3)	(4)	(5)
Peer Evaluator	-0.510	-0.728	-0.846	-0.683	-0.728
	(0.376)	(0.143)	(0.149)	(0.189)	(0.143)
Multiple-Prize	-1.567***	-1.504***	-1.215**	-1.627***	-1.504***
	(0.007)	(0.003)	(0.042)	(0.002)	(0.003)
Multiple-Prize × Peer	-0.972*	-1.499***	-0.464	-1.842***	-1.499***
	(0.095)	(0.003)	(0.437)	(0.000)	(0.003)
Core Evaluators	✓				
All Evaluators		✓			
Manager Evaluators			✓		
Peer Evaluators				✓	
Assigned Evaluators					✓
Demographic Controls	✓	✓	✓	✓	✓
Adj. R-Square	0.053	0.067	0.054	0.070	0.067
Observations	527	527	527	527	527

Notes. The specifications in this table mirror those in Column (1) of Table 5, with Column (1) being an exact replication, but with different score aggregation approaches for the dependent variables. All regression specifications include a constant, but we do not report the constant for brevity. We estimate heteroscedasticity-consistent standard errors and report two-tailed *p*-values in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively

Table B.2: *Effect of Design Choices on Participation Behavior (Outlier Adjusted)*
(two-tailed p-values in parentheses)

Panel A: Removing Outlier

Dependent Variable:	Number of Ideas Submitted	Submitted at Least One Idea
	(1)	(2)
Peer Evaluator	0.077*	0.009
	(0.088)	(0.521)
Multiple-Prize	0.060	0.020
	(0.194)	(0.140)
Multiple-Prize × Peer	0.064	0.016
	(0.154)	(0.246)
Demographic Controls	✓	✓
Adj. R-Square	0.012	0.036
Observations	940	940

Panel B: Adjust Outlier Value

Dependent Variable:	Number of Ideas Submitted	Submitted at Least One Idea
	(1)	(2)
Peer Evaluator	0.092*	0.010
	(0.054)	(0.481)
Multiple-Prize	0.043	0.019
	(0.377)	(0.161)
Multiple-Prize × Peer	0.049	0.015
	(0.302)	(0.274)
Demographic Controls	✓	✓
Adj. R-Square	0.007	0.035
Observations	941	941

Notes. These results mimic those in Table 3, albeit we make adjustments to deal with outliers. In Panel A, we remove one outlier from the sample, who submitted 26 ideas. In Panel B, we replace the value of the discrete participation variable in Column (1)—the number of ideas submitted—to be the second highest value, 15, for this individual. The results are similar if we set the upper bound of the discrete participation variable to be 10, 11, 12, 13, 14, or 15. We estimate heteroscedasticity-consistent standard errors and report two-tailed *p*-values in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

C. Most Useful, Novel, and Creative Ideas

Panel A: Manager Evaluators and \$100×1

Usefulness	Score	
1	70.6	Another idea is that of a mini-contest. Although I have never won a contest on Mturk, I do believe my fellow participants like having a chance at winning a prize (that is part of the reason why we are here). If a participant passes the attention checks, not only are they able to continue with the survey, but they are also qualified to win a prize, with most participants getting a bonus of some sort. Participants love bonuses, no matter what these bonuses may be.
2	68.7	Requester might try attaching a small attention bonus. On a page where a worker needs to be attentive, they can have a word flash on the screen for a couple of seconds. If the worker records the word, they get a little extra monetary compensation added on to the base survey payment. Maybe there could be 3 or 4 so the worker never knows when it's coming. If they catch all of them, then maybe there could be an a slightly larger bonus. Maybe the words could form a phrase that the worker has to enter at the end to unlock the bonus.
3	68.1	How about a sentence that you would be asked to take the first letter form each work to spell a day of the week. For instance Frank ran in Davids alley yesterday, that would be FRIDAY.
Novelty	Score	
1	77.6	Intergalactic Explorer – The participant chooses an avatar and enters a spaceship; they direct their spaceship from galaxy to galaxy, picking up lifeforms, relevant space objects, etc. that are needed at their home planet. The objects/lifeforms they pick up may be associated/labelled with items/facts they have read in the survey up to that point. Also, on their journey, when they find a necessary object and thus demonstrate comprehension, they are given a clue to find a special mystery treasure in that galaxy. If they find it, they receive a small bonus. By the end of the journey, the turker should have collected a prescribed amount of objects/correct answers to pre-given questions. At that point, they can return home, count of their treasure, and continue with the survey.
2	76.5	Using the scene from the Wizard of Oz, where the trees hurl their apples at the travelers (as they cry 'how bout dem apples'), a scary trees throws apples at the worker's avatar, which is perhaps a character from the famous movie or one chosen by the worker. The apples are labelled with facts, ideas, answers from the survey up to that point, as well as false items. The worker must evade the 'false' apples and catch the apples with the correct items. If they are successful, they move on down the 'yellow brick road.' If not, the flying monkeys come and carry them away.

3	76.4	'The M-Turker 500' – The survey taker chooses from a selection or is given a racecar. They will race against a number of other computer cars. The challenge is that their car slows down successively during the race if they do not answer comprehension questions accurately and in a timely manner. On the other hand, the quicker they answer questions more accurately, the quicker their car goes. Also, their car may also blow a gasket, get a flat tire, or run out of fuel at random. These calamities can only be fixed by responding to comprehension-related questions when they flash across the screen as quickly as possible. The race lasts for a prescribed amount of time, say 4 min. (unless the racer gains time through quick, accurate answers and finishes before that time). If the racer finishes and wins before the 4 minute mark, they get a bonus. If they they are in the lead but do not finish the laps at the end of 4 min., they get a smaller bonus. If they get second place, they can continue with the survey. Third or lower and they may be dropped or have to redo earlier sections of the survey.
Creativity	Score	
1	71.0	Intergalactic Explorer – The participant chooses an avatar and enters a spaceship; they direct their spaceship from galaxy to galaxy, picking up lifeforms, relevant space objects, etc. that are needed at their home planet. The objects/lifeforms they pick up may be associated/labelled with items/facts they have read in the survey up to that point. Also, on their journey, when they find a necessary object and thus demonstrate comprehension, they are given a clue to find a special mystery treasure in that galaxy. If they find it, they receive a small bonus. By the end of the journey, the turker should have collected a prescribed amount of objects/correct answers to pre-given questions. At that point, they can return home, count of their treasure, and continue with the survey.
2	70.7	Have mTurkers do a flower arranging task. Invite them to follow specific directions for arranging a bouquet of flowers. To the side is a box with a variety of flowers, in the middle a vase. By clicking an dragging, they must choose the correct flower and place each one in the correct position in the vase. The benefit of this there is nothing tricky about it, it is straight forward, but it has some aesthetic appeal (especially to those who like flowers) if the images of the flowers and vase are appealing enough.
3	70.3	How about a sentence that you would be asked to take the first letter form each work to spell a day of the week. For instance Frank ran in Davids alley yesterday, that would be FRIDAY.

Panel B: Manager Evaluators & \$10×10

Usefulness	Score

1	71.1	Rather than using attention checks as a rejection threat, turn them into a challenge. Add 5 checks throughout the survey, and inform turkers that if they can find them and correctly answer them, they will receive a bonus. This way people will pay a lot more attention to the survey because they will be carefully looking for the checks. This method works much better than the threat of rejection, and people will work much more carefully.
2	70.3	Easy, use a CAPTCHA but tell them to get it wrong. It's something that people who are on autopilot absolutely won't catch, but anyone paying the slightest bit of attention will catch.
3	68.6	Have the last page of the survey display 10 questions and ask the participant to identify which of the questions had been asked, and what their answer was. Give a bonus for each right answer.
Novelty	Score	
1	75.3	I would add a little turtle traveling across the bottom. He arrives at different places (school, gas station, post office)as the survey progresses. The attention check could be to describe where the turtle is on his journey.
2	72.4	Easy, use a CAPTCHA but tell them to get it wrong. It's something that people who are on autopilot absolutely won't catch, but anyone paying the slightest bit of attention will catch.
3	71.2	Here's another idea- the #1 pet peeve most MTurkers would agree on is the fact that many requesters forget to (or intentionally) don't provide completion codes... set up your attention checks so that they give the MTurk worker a segment of the code as they go. Not only does it reward attention, it also assures that the survey is not a scam...
Creativity	Score	
1	72.3	I would add a little turtle traveling across the bottom. He arrives at different places (school, gas station, post office)as the survey progresses. The attention check could be to describe where the turtle is on his journey.
2	66.5	Easy, use a CAPTCHA but tell them to get it wrong. It's something that people who are on autopilot absolutely won't catch, but anyone paying the slightest bit of attention will catch.
3	65.6	Have a little treasure chest picture (or money bag) on the survey that is interactive. When there are things that need to be paid attention to, the picture could glow or be highlighted, or some other way to see that it needs to be clicked on. By clicking on it the person will be asked a couple of comprehension questions, which will result in a small bonus for every question they get right. By seeing the treasure chest/money bag even if it's not active, people would probably be more engaged since there's an obvious additional reward if they are attentive.

Panel C: Peer Evaluators & \$100×1

Usefulness	Score	

1	76.2	most of the attention checks are dull and unoriginal; others are so tricky that they defeat the whole purpose of an attention check which is to make sure people are paying attention; not to trick them. my idea is to put in simple arithmetic questions; questions that any 3rd grader could answer but that are not tricky. for example: fill in the blank: 1, 3, (?), 7, 9. this requires attention but is easy. obviously the answer is 5.
2	73.4	I think that it works well when questions are thrown in the mix that have nothing to do with the survey but occur on the same page as valid questions. Often, surveys have a scale of Totally Disagree (for instance) all the way to Totally Agree. Statements like..,ÄôI have been mortally wounded by a bear,,Äù should only merit a Totally Disagree answer. Anything other than that needs to be thrown out. Entire pages that are dedicated to an attention check are not optimal. If there is a random page with a multiple choice of colors, you know the question is an attention check and often no additional checks occur after that. The checks need to be sprinkled throughout.
3	69.4	At the beginning of a survey, mention a secret code word or number hidden somewhere in the survey that must be entered at some point during the experiment. The secret code could be placed in text that workers have to read, ensuring instructions or scenarios are read thoroughly to obtain the secret word. It doesn't have to be anything difficult, Something easily remembered mentally. Make it known that the attention check could pop up at any moment and in order to proceed with the HIT and have work approved, you must answer correctly with the code word provided. I feel like workers that are actually paying attention would be motivated and anticipate the attention check. We'd be more apt to read through everything on the screen to find the code instead of the standard long paragraph that we usually get that says, "In order to demonstrate you read the instructions...click this and type this."

Novelty	Score	
1	75.4	Alchemy. A chart will show various combinations of colored liquids and the alchemical result of the mix. Below the chart, there will be 5 empty vials. Each click on a vial will change it's color to the next option. Using the chart you must find the correct combination of liquids which is called for such as green green red blue purple. The vials must be changed to match this in order to proceed.
2	73.2	Color in the black cat to proceed. Cat is divided into segments. Multiple color options will be available but the correct one that should be used is black for every segment.

3	72.2	For this idea take a picture of magazine that represents the survey the worker is working on with all kinds of articles or pictures in it. Once the worker gets to this part of the survey where they have either lost interest or not paying attention is magazine will appear on the screen. As the worker either clicks the enter button or space bar the magazine will flip between pages. It will either read about make-up blog, vacation home, food craving, new products that are coming or etc. It will last for about five minutes or depending on how long the survey is. When they get to the end of it a roller coaster will pop up with a sweet smiling panda that says, "HAHA your IT." Then the panda will start dancing to music as he or she slowly drifts away from the screen. The worker will get an arrow to proceed because they actually pay attention within the time frame prohibited or an exit sign to not return the survey.

Creativity	Score	
1	67.7	Alchemy. A chart will show various combinations of colored liquids and the alchemical result of the mix. Below the chart, there will be 5 empty vials. Each click on a vial will change it's color to the next option. Using the chart you must find the correct combination of liquids which is called for such as green green red blue purple. The vials must be changed to match this in order to proceed.
2	67.4	An attention check should be straightforward and not sneaky. If you try to do something too tricky, people will be upset. A fun and easy to do attention check would be to present a cartoon with no writing and have people create a caption for it. Making people be creative would help them to better focus on the rest of your survey while being fun at the same time.
3	66.4	This idea is regarding the character Garfield. Garfield is flying an airplane in this video with one hand on the wheel and with the other hand with thumbs up. He is flying a blue airplane with the letters written in red and white that reads flight 47. He also has have sunglasses and of course a black helmet with a white scarf wrap around his neck. There is a hot pink flyer at the end of the airplane reading in bold yellow letters,"Will be back shortly as survey turns." For every worker that either is not paying attention this can pop up as they get to the point of the survey in which it flash across the screen to catch their attention.

Panel D: Peer Evaluators & \$10×10

Usefulness	Score	
1	68.1	Have them pick out the odd thing out in an image of four things - which one of these things is not like the others? As an example, show a chocolate chip cookie, a slice of cake, an ice cream cone, and portion of kale.

2	67.5	I think the best way to ensure that mturkers pay attention is to provide questions that the users have to actually read, but are absurd and only have one correct answer. For instance.. I enjoy riding buffalos in the morning. If included in the survey they aren't easy to spot unless you are actually reading the question and have a clear answer. Some other examples of possible questions - Riding dinosaurs is a hobby I enjoy– answer is true or false - I have gills and swim in the ocean– answer yes or no
3	66.6	To gain interest make it fun, use trivia questions as attention checks, when the Turker gets it right they get a nickel/dime...whatever, even a couple of pennies...BUT embed the questions in with your scales/etc. make sure they have to read them... like "I feel depressed often...the first US president was Washington.. Agree, disagree, etc.Make the total amount of hit something like .50 or .75 but put in TITLE...ALL CAPS OR big..font, Fun Trivia quiz, win 2.00 in bonus...or whatever. Turkers love Trivia stuff, they also love things that can be sort of a game or easter egg hunt, it makes it interesting and it will be to the turkers GREAT advantage to pay attention. Warning though, make each page timed in some way so they can't really look things up but DON'T make it so tough that they can't figure it out.... Maybe make it a history or geography theme? You pick, even classic TV or Academy award winning movies, have fun with it.

Novelty	Score	
1	68.0	An interesting attention check would be to have the worker properly complete a mad lib. It seems like it would be easy enough to code depending on the variations in input desired.
2	67.1	Play a flash game of Pong as an attention check against a less than average AI. If you beat the AI, not only do you pass the attention check, you also receive a good bonus such as a dollar. 90% of the reason people don't pass attention checks is because the typical pay for the study is ludicrous. 12/hour should be standard.
3	66.7	The attention check should be a picture of an acrostic poem, which is the kind of poem where the first letter of every sentence/phrase/word forms a different word. The attention check should be a text box where someone has to identify the word that forms the basis of the acrostic poem.

Creativity	Score	
1	66.1	Create a game or detective game within the survey. Leave clues throughout that are in bold or the reader has to find themselves. These clues will stand out because they have nothing to do with the survey and they would know it was a clue for that reason and would have to write it down. In the end they would have to provide the answer or list all clues and come up with an answer to all the clues. Keep it fun and if they don't get the answer right it would be ok and not get rejected because they had all the clues and that would be the main thing to make sure they were paying attention. The person that gets the answer correct should get an extra bonus of your discretion. Make sure to leave extra time for the hit because some will take longer than others.

2	65.2	Have the participant choose a word that does NOT follow a common grammar rule (Which of the following does not follow the rule ‘i before e except after c’? 1. Receipt, 2. Deceive, 3. Weird, 4. Tier)
3	64.1	The attention check should be a picture of an acrostic poem, which is the kind of poem where the first letter of every sentence/phrase/word forms a different word. The attention check should be a text box where someone has to identify the word that forms the basis of the acrostic poem.

D. Evaluator Preferences for Usefulness and Novelty

As we discussed in the main text, creative ideas are those that are both novel and useful [Amabile 1983]. When ideas are evaluated subjectively, the evaluator must decide how much weight to place on each feature when making a creativity assessment. We argued that an evaluator's role within an organization is likely to influence the weight they place on novelty and usefulness. We posit that manager evaluators place greater emphasis on usefulness as a driver of creativity than do peer evaluators, because managers have to bear the costs of implementing the ideas. As such, they will place additional value on realistic ideas with a clear benefit to the firm. As discussed in the main text, we predict and find evidence in support of the belief that participants anticipate this and adjust their submissions accordingly.

Here we empirically test whether managers actually place more emphasis on usefulness by regressing an idea's creativity score on its novelty and usefulness scores. We do this in two separate regressions, with the first considering the scores given by peer evaluators and the second considering the scores given by manager evaluators. We then compare the relation between peer usefulness scores and peer creativity scores to the relation between manager usefulness scores and manager creativity scores.

Panel A of Table D.1 considers the scores given by peer evaluators. In Column (1), we use the main sample of unique ideas and regress the average peer creativity score given to each idea on the average peer usefulness score and average peer novelty score given to the idea. In Column (2), instead of using average scores for each idea, we expand the sample to include each individual score given to each idea. There were 8,961 unique idea-evaluator scores given by peer evaluators and 3,314 unique idea-evaluator scores given by manager evaluators. In Column (3) we include evaluator fixed effects, and in Column (4) we include idea fixed effects to ensure our results are not heavily influenced by the scores given by a single evaluator or by the scores received by a single idea. Panel B is analogous to Panel A, but it considers the scores given by manager evaluators.

Table D.1 shows that usefulness and novelty each relate positively and significantly to creativity among *both* peer and manager evaluators. To test the notion that manager evaluators emphasize usefulness more than do peer evaluators, we compare the magnitude of the point estimates on *Peer Usefulness Score* and *Manager Usefulness Score*. The coefficients on *Manager Usefulness Score* are 2.7–4.2 times larger than the coefficients on *Peer Usefulness Score*. The differences between the coefficients are statistically significant at the 1% level across all columns, suggesting that manager evaluators place significantly more weight on usefulness as a feature of creativity than do peer evaluators. The differences between the point estimates on *Peer Novelty Score* and *Manager Novelty Score* are also statistically significant, suggesting that peer evaluators place more weight on idea novelty when judging the creativity of an idea than do manager evaluators. This empirical evidence aligns with our earlier propositions and supports our idea creativity results discussed in Section 4.2.

Table D.1: *Creativity as a Function of Novelty and Usefulness*
(two-tailed p-values in parentheses)

Panel A: Peer Evaluations

	Peer Creativity Score			
Peer Usefulness Score	0.086***	0.141***	0.110***	0.127***
	(0.011)	(0.007)	(0.006)	(0.007)
Peer Novelty Score	0.947***	0.863***	0.826***	0.776***
	(0.011)	(0.006)	(0.006)	(0.008)
Evaluator Fixed Effect			✓	✓
Idea Fixed Effect				✓
Adj. R-Square	0.953	0.838	0.875	0.880
Observations	527	8,961	8,961	8,961

Panel B: Manager Evaluations

	Manager Creativity Score			
Manager Usefulness Score	0.361***	0.378***	0.373***	0.401***
	(0.017)	(0.010)	(0.010)	(0.011)
Manager Novelty Score	0.716***	0.630***	0.613***	0.523***
	(0.013)	(0.010)	(0.010)	(0.014)
Evaluator Fixed Effect			✓	✓
Idea Fixed Effect				✓
Adj. R-Square	0.926	0.828	0.838	0.845
Observations	527	3,314	3,314	3,314

Notes. These results are meant to compare the relation between peer usefulness (novelty) scores and peer creativity scores to the relation between manager usefulness (novelty) scores and manager creativity scores. Panel A considers the scores given by peer evaluators, and Panel B considers the scores given by manager evaluators. In Column (1), we use the main sample of unique ideas and regress the average peer creativity score given to each idea on the average peer usefulness score and average peer novelty score given to each idea. In Column (2), instead of using average scores for each idea, we expand the sample to include each individual score given to each idea. There were 8,961 unique idea-evaluator scores given by peer evaluators and 3,314 unique idea-evaluator scores given by manager evaluators. In Column (3) we include evaluator fixed effects, and in Column (4) we include idea fixed effects. Regression specifications include a constant, but we do not report the constant for brevity. We estimate heteroscedasticity-consistent standard errors and report two-tailed *p*-values in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively