NBER WORKING PAPER SERIES

CREATIVITY UNDER FIRE:
THE EFFECTS OF COMPETITION ON CREATIVE PRODUCTION

Daniel P. Gross

Working Paper 25057
http://www.nber.org/papers/w25057

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2018

I am grateful to Ben Handel, Steve Tadelis, and especially John Morgan, whose conversation provided much of the inspiration for this paper. I also thank Gustavo Manso and Noam Yuchtman for feedback at early stages, colleagues at UC Berkeley and Harvard Business School, and numerous seminar and conference audiences for their comments. Limited short segments of the text may be similar to passages from another of my papers ("Performance Feedback in Competitive Product Development"), which uses a different dataset from the same setting. This research was supported by NSF Graduate Research Fellowship Grant No. DGE-1106400 as well as the Harvard Business School Division of Faculty and Research. All errors are my own. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Creativity Under Fire: The Effects of Competition on Creative Production
Daniel P. Gross
NBER Working Paper No. 25057
September 2018
JEL No. D81,M52,M55,O31,O32

ABSTRACT

Though fundamental to innovation and essential to many industries and occupations, individual creativity has received limited attention as an economic behavior and has historically proven difficult to study. This paper studies the incentive effects of competition on individuals' creative production. Using a sample of commercial logo design competitions, and a novel, content-based measure of originality, I find that intensifying competition induces agents to produce original, untested ideas over tweaking their earlier work, but heavy competition drives them to stop investing altogether. The results yield lessons for the management of creative workers and for the implementation of competitive procurement mechanisms for innovation.

Daniel P. Gross
Harvard Business School
Soldiers Field
Boston, MA 02163
and NBER
dgross@hbs.edu

The creative act is a broadly important but under-studied phenomenon in economics. Millions of people in the U.S. alone work in fields where creativity is essential to job performance, such as research, engineering, and professional services – industries which are the engines of innovation and growth in modern developed economies. CEO surveys also show that executives' top concerns consistently include the creativity of their employees and pursuit of innovation within the firm. Despite its importance, the creative act itself has received limited attention as an economic behavior and has historically proven difficult to study, due to the challenge of measuring creativity and relating it to variation in incentives.

This paper studies the incentive effects of competition on individuals' creative output, exploiting a unique field setting where creative activity and competition can be precisely measured and related: tournaments for the design of commercial logos and branding. Using image comparison tools to measure originality, I show that intensifying competition both creates and destroys incentives for creativity. While some competition is necessary to induce high-performing agents to develop original, untested designs over tweaking their existing work, heavy competition discourages effort of either kind. Theory suggests these patterns are driven by the risk-return tradeoffs inherent to innovation. In the data, agents are most likely to produce original designs in a horserace against exactly one other competitor of similar quality.

It is useful to begin with a definition: creativity is the act of producing ideas that are novel and appropriate to the goal at hand (Amabile 1996, Sternberg 2008). The paper opens with a simple model that provides a framework for thinking about the economics of creative activity in a tournament setting, which both guides the empirical analysis and rationalizes its results.[1] In this model, a principal seeks a new product design and solicits candidates from a pool of workers via a tournament, awarding a prize to the best entry. Workers enter designs in turns, and once entered, each submission's quality is public knowledge. At each turn, workers must choose between developing an original design or tweaking a previous entry, cast as a choice between an uncertain and safe outcome. The model suggests that competition increases workers' incentives to produce original designs over tweaks – but it also shows that heavy competition depresses incentives to do either. Though intuitive, and in part a recasting of prior theoretical research to this paper's context, the model is useful in framing and interpreting empirical results throughout the paper.

The paper then turns to an empirical study of logo design competitions, drawing on a sample of contests from a popular online platform.[2] In these contests, a firm ("sponsor") solicits custom designs from freelance designers ("players"), who compete for a winner-take-all prize. The contests in the sample offer prizes of a few hundred dollars and on average attract around 35 players and 100 designs. An important feature of

[1] The model in this paper is related to Taylor (1995), Che and Gale (2003), Fullerton and McAfee (1999), and Terwiesch and Xu (2008) but differs in that it injects an explore-exploit dilemma into the agents' choice set: whereas existing work models competing agents who must choose how much effort to exert, the agents in this paper must choose whether to build off of an old idea or try a new one, much like a choice between incremental versus radical innovation. The framework also has ties to recent work on tournaments with feedback (e.g., Ederer 2010), bandit problems in single-agent settings (Manso 2011), and models of competing firms' choice over R&D project risk (Cabral 2003, Anderson and Cabral 2007).

[2] The empirical setting is conceptually similar to coding competitions studied by Boudreau et al. (2011), Boudreau et al. (2016), and Boudreau and Lakhani (2015), though the opportunity to measure originality is unique. Wooten and Ulrich (2013, 2014) have also studied graphic design competitions, focusing on the effects of visibility and feedback.

this setting is that the sponsor can provide real-time feedback on players' designs in the form of 1- to 5-star ratings. These ratings allow players to gauge the quality of their own work and the intensity of competition while the contest is underway. Most importantly, the dataset also includes the designs themselves, which makes it possible to study creative choices over the course of a contest: I use image comparison algorithms similar to those used by commercial content-based image retrieval software (e.g., Google Image Search) to calculate similarity scores between pairs of images in a contest, which I then use to quantify the originality of each design relative to prior submissions by the same player and her competitors.

This setting presents a unique opportunity to observe creative production in the field. Though commercial advertising is important in its own right, the iterative product development process observed here is similar to that in other domains where prototypes are created, tested, and refined. The nature of the setting enables a more detailed empirical study of this process, and its interaction with incentives, than is typically possible. The tournament format is especially germane: although the website advertises itself as a crowdsourcing platform, the contracting environment is fundamentally a request for proposals (RFP), a competitive mechanism widely used by firms and government agencies to procure new products or technologies – often over multiple rounds, with interim scoring, and typically with only the top bid rewarded.

The sponsors' ratings are critical in this paper as a source of variation in the information that both I and the players have about the state of the competition. Using these ratings, I am able to directly estimate a player's probability of winning, and the results establish that ratings are meaningful: the highest-rated design in a contest may not always win, but a five-star design increases a player's win probability as much as 10 four-star designs, 100 three-star designs, and nearly 2,000 one-star designs. Data on the time at which designs are entered by players and rated by sponsors makes it possible to establish what every participant knows at each point in time – and what they have yet to find out. The empirical strategy exploits naturally-occurring, quasi-random variation in the timing of sponsors' ratings and compares players' responses to information they observe at the time of design against that which is absent or not yet provided.

I find that competition has large effects on the content of players' submissions. Absent competition, positive feedback causes players to cut back sharply on originality: players with the top rating produce designs more than twice as similar to their previous entries than those with only low ratings. The effect is strongest when a player receives her first five-star rating – her next design will be a near replica of the highly-rated design – and attenuates at each rung down the ratings ladder. However, these effects are reversed by half or more when high-quality competition is present: competitive pressure counteracts this positive feedback, inducing players to produce more original designs. A battery of supporting analysis establishes that this result is econometrically identified and is robust to alternative measures of the key variables.

Taken alone, these results suggest that competition unambiguously motivates creativity, but the analysis, and conclusion, presumes no outside option. In practice, players have a third option: they can stop bidding. Whether and when this alternative becomes binding is its own question. Consistent with previous research (e.g., Baik 1994 or Brown 2011), I find that heavy competition discourages further investment. Empirically, high performers' tendency to produce original work is greatest when facing roughly 50-50 odds of winning – in other words, when neck-and-neck against one similar-quality competitor.

The driving assumption behind the model, and the interpretation of these results, is that creative effort is risky but high-return. The data indicate that original designs outperform tweaks of low-rated work, but due to the ratings being bounded above at five stars, the same cannot be observed against tweaks of high-rated work. To test this assumption, I recruit a panel of professional designers to administer independent ratings of five-star designs on an extended scale and correlate their responses with these designs' originality. I find that original designs are on average more highly-rated by these panelists than tweaks, but the distribution of opinion also has higher variance, reflecting risk. This evidence thus reinforces a possible link between creativity and risk-taking which has been suggested by research in other fields.

These findings contribute to a developing but mixed literature on the effects of competition on individual creative output: economists argue that competition can motivate the kind of risk-taking that is characteristic of inventive activity (e.g. Cabral 2003, Anderson and Cabral 2007), yet many psychologists argue that high-powered incentives and other extrinsic pressures stifle creativity by crowding out intrinsic motivation (see Amabile and Hennessey 2010 for a review) or by causing agents to choke (Ariely et al. 2009). Lab-based studies of creativity are as mixed as the theory (e.g., Eisenberger and Rhoades 2001, Charness and Grieco 2018, Erat and Gneezy 2016, Bradler et al. 2018), in part due to differences in measurement and experimental design. Missing from the creativity literature is the added nuance that competition is not strictly a binary condition but rather can vary in intensity across treatments – and as this paper shows, the effects hinge crucially on the intensity of competition, as well as the existence of an outside option.

The evidence that creativity can be elicited with balanced competition has substantive implications for managers in creative industries and for the procurement practices of all organizations. Many practitioners appear to subscribe to the aforementioned intrinsic motivation theory of creativity endorsed by social psychologists, which holds that extrinsic motivators are counterproductive and is regularly communicated in the Harvard Business Review (e.g., Florida and Goodnight 2005, Amabile and Khaire 2008, Amabile and Kramer 2012) and other business press. While intrinsic motivation is valuable, the results of this paper demonstrate that high-powered incentives can be effective at motivating creativity, if properly managed. The results also provide lessons for organizers of innovation prize competitions and other competitive procurement mechanisms for innovation (e.g., RFPs) on managing the intensity of competition.

The paper also makes a methodological contribution to the innovation literature. Due to data constraints, empirical research has historically measured innovation in terms of inputs (such as R&D spending) or outputs (patents), when innovation is at heart about the individual acts of discovery and invention that take place between. As a result, there is relatively little systematic, empirical evidence on the process of idea production. This paper is an effort to fill this gap, invoking new tools for content-based measurement of innovation and using them to study how ideas are developed and refined in response to incentives.

The paper is organized as follows. Section 1 discusses related literatures in economics and social psychology and presents the model. Section 2 introduces the empirical setting and describes the identification strategy. Section 3 estimates the effects of competition on submissions' originality. Section 4 presents the countervailing effects on participation. Section 5 provides evidence that creativity is risky but high-return, supporting the key assumption of the model. Section 6 discusses the implications of these results for policy, management, and future research on creativity and innovation and concludes.

1 Background: Creativity and Incentives

1.1 Existing Literature

Research on individual creativity has historically belonged to the realm of social psychology. The question of whether incentives enhance or impair creativity is itself the focus of a contentious, decades-old debate led by two schools of thought: one camp argues that incentives impair creativity by crowding out intrinsic motivation (Amabile 1996, Hennessey and Amabile 2010), whereas the other argues that incentives bolster creativity, provided that creativity is explicitly what is being rewarded (Eisenberger and Cameron 1996). Scholars in each of these camps have written public rejoinders to the other (e.g., Eisenberger and Cameron 1998, Hennessey and Amabile 1998), while others have sought to develop and test more nuanced theories in an attempt to reconcile these arguments (e.g., Shalley and Oldham 1997).

The empirical literature on which these arguments are based in most cases invokes high-powered incentives (tournaments) in its experimental design. Despite dozens of experiments, the empirical evidence has been unable to clarify which of these positions is valid (Shalley et al. 2004). Different papers include different-sized rewards (which may or may not not be valuable enough to overcome motivational crowd-out, to the extent it occurs), different subject pools (college students versus grade-school children), and inconsistencies in how performance is evaluated and what features of performance are rewarded: studies cited by the pro-incentives camp reward subjects for creativity, whereas studies cited by the anti-incentives camp evaluate creativity but often reward the best ideas. Experiments on both sides rely heavily on judges' assessments of creativity, which they are typically asked to score according to their own definitions.

Experimental economists have recently entered the literature, though often subject to the same limitations. Erat and Gneezy (2016) evaluate subjects' creativity in a puzzle-making task under piece-rate and competitive incentives and find that competition reduces creativity relative to an incentive-free baseline. Charness and Grieco (2018) in contrast find that high-powered incentives increase creativity in closed-ended creative tasks and have no effect on creativity in open-ended tasks. In both studies, creativity is scored by judges without guidance or a standardized definition, which leads to low inter-rater reliability. Rather than relying on subjective assessments, Bradler et al. (2018) study the effects of tournament incentives and gift exchange on creative output with an unusual uses task – where subjects are asked to think of productive uses for a common household object (e.g., a tin can), and creativity is measured by the statistical infrequency of each answer. In this case, the authors find that tournaments increase creative output relative to both gift exchange and an incentive-free baseline, though the empirical methodology makes it hard to distinguish an increase in originality (novel uses) from an increase in output alone (total uses).

This paper makes several important departures from this body of research. The logo design competitions studied here provide a field setting in which real creative professionals are competing for prizes of significantly greater value than observed in the existing lab-based studies. They also provide a setting where originality can be objectively measured with content-based assessment. Additionally, in contrast to much of the literature, it is not creativity per se that is being rewarded, but rather product quality: as in most product development settings, creativity here is a means towards an end, rather than an end in and of itself. Most importantly, however, this paper studies competition as a continuously-varying rather than binary treatment. In practice, competition is not a uniform condition, and the fact that the implementation of competitive incentives varies across the previously-cited studies might perhaps even explain their divergence.

At the heart of this paper is a set of empirical results on the originality of submissions into the sampled tournaments. To the extent that being creative is risky and its outcome uncertain, as the model below will propose, the paper is also connected to the economics literature on choices over risk in competition. Cabral (2003) and Anderson and Cabral (2007) show that in theory, laggards will take actions with higher-variance outcomes, consistent with the intuition of needing a big hit to catch up or, in the extreme, of having nothing to lose. Similar behavior has been observed among investment fund managers, who increase fund volatility after a mid-year review which reveals them trailing their peers' performance (Brown et al. 1996) or when trailing the market (Chevalier and Ellsion 1997). In additional related work, Genakos and Pagliero (2012) study the choice over how much weight to attempt across rounds of dynamic weightlifting tournaments, and interpret the decision as a choice over risk. The authors find that whereas moderate laggards increase risk, distant laggards reduce risk – a result at odds with the existing theoretical literature and the evidence from this paper, which indicate that more distant laggards prefer greater risk, conditional on participating. The interpretation, however, may be limited by the difficulty of empirically distinguishing a choice of risk from a commitment to a specified level of effort in the weightlifting context.

An additional literature to which this paper relates is the long-running literature in economics on product market competition and innovation (see Gilbert 2006 and Cohen 2010 for summaries). Since Schumpeter's (1942) contention that market power is favorable to innovation, researchers have produced explanations for and evidence of positive, negative, and inverted-U relationships between competition and innovation in a variety of markets – though the literature is complicated by differences in definition and measurement, challenges in econometric identification, and institutional variation. In a seminal contribution, Aghion et al.

(2005) predict an inverted-U effect of product market competition on step-by-step innovation, and Aghion et al. (2014) find support for the predictions of this model in a lab experiment designed to mimic its features. There are, however, a few key differences between this paper's setting and the Aghion et al. (2005) model, the most important of which are the emphasis on individual creative behavior and the tournament context, where innovation is continuous and the intensity of competition is determined by relative performance differences, rather than by an exogenous degree of collusion in the product market.

1.2 Theoretical Framework

The preceding literature explains creativity and its motives primarily through narrative psychological constructs, rather than economic forces. Yet creativity can also be interpreted as an economic behavior, insofar as it involves a choice over uncertainty. This section demonstrates how this idea can be operationalized in a relatively simple tournament model whose features resemble the empirical setting. The results are presented in partial equilibrium to bring into focus the tradeoffs facing agents in such a setting, which both guide the empirical analysis and offer a framework for interpreting the evidence that follows.

Suppose a risk-neutral principal seeks a product design. Because R&D outcomes are uncertain and difficult to value, the principal cannot contract directly on performance. It instead sponsors a tournament to solicit prototypes from J risk-neutral players, who enter designs sequentially and immediately learn of their quality. Each design can be either original or adapted from the blueprints of previous entries; players who choose to continue working on a given design at their next turn can re-use the blueprint to create variants, with the original version remaining in contention. At each turn, the player must decide whether to continue investing and if so, whether to produce an original design or tweak an earlier submission. At the end of the tournament, the sponsor awards a winner-take-all prize P to its favorite entry.

Let each design be characterized by latent value ν_{jt}, which only the sponsor observes:

$$\nu_{jt} = \ln\left(\beta_{jt}\right) + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim \text{i.i.d. Type-I E.V.} \tag{1}$$

where j indexes players and t indexes designs. In this model, β_{jt} represents the design's quality, which is revealed by the sponsor's feedback, and the latent value is a function of revealed quality and a i.i.d. random shock, which reflects idiosyncracies in the winner selection process. To hone intuition, further suppose each player enters at most two designs. The type-I extreme value error leads to logit choice probabilities for each design (see Train 2009), such that player j's total probability of winning is:

$$Pr\left(\text{player } j \text{ wins}\right) = \frac{\beta_{j0} + \beta_{j1}}{\beta_{j0} + \beta_{j1} + \sum_{k \neq j}\left(\beta_{k0} + \beta_{k1}\right)} = \frac{\beta_{j0} + \beta_{j1}}{\beta_{j0} + \beta_{j1} + \mu_j} \tag{2}$$

where $\mu_j \equiv \sum_{k \neq j}\left(\beta_{k0} + \beta_{k1}\right)$ is the competition that player j faces in the contest. This function is concave in the player's own quality and decreasing in the quality of her competition.

Every player's first design in the contest is inherently novel, and entry is taken for granted – in theoretical terms, each player is endowed with their first submission. At their subsequent turn, they have three options: they can exploit (tweak, or adapt) the existing design, explore (experiment with) a radically different design, or abandon the contest altogether. To elaborate on each option:

1. **Exploitation** costs $c > 0$ and yields a design of the same quality, resulting in a second-round design with $\beta_{j1} = \beta_{j0}$ and increasing the player's probability of winning accordingly.

2. **Exploration** costs $d \geq c$ and can yield a high- or low-quality design. With probability q, exploration will yield a high-quality design with $\beta_{j1}^{H} = \alpha\beta_{j0}$; with probability $(1-q)$ it will yield a low-quality design with $\beta_{j1}^{L} = \frac{1}{\alpha}\beta_{j0}$, where $\alpha \geq 1$ is the exogenous degree of exploration.[3]

3. **Abandonment** is costless: the player can abstain from further investment. Doing so leaves the player's probability of winning unchanged, as her previous work remains in contention.

In this context, feedback has three effects: it informs each player about her first design's quality, influences her second design, and reveals the level of competition she faces. Players use this information to decide (i) whether to continue participating and (ii) whether to do so by exploring a new design or re-using a previous one, which is a choice over which kind of effort to exert: creative or rote.

Conditions for Exploration

To further simplify notation, let $F(\beta_1) = F(\beta_1 | \beta_0, \mu)$ denote a player's probability of winning when her second submission has quality β_1, given an initial submission of quality β_0 and competition μ (omitting the j subscript). For a player to produce an original design, she must prefer doing so over both exploiting the existing design (Eq. 3.1) and abandonment (Eq. 4.1):

$$\underbrace{\left[qF\left(\beta_1^H\right) + (1-q)\,F\left(\beta_1^L\right)\right] \cdot P - d}_{E[\pi|\text{explore}]} > \underbrace{F\left(\beta_0\right) \cdot P - c}_{E[\pi|\text{exploit}]} \tag{3.1}$$

$$\underbrace{\left[qF\left(\beta_1^H\right) + (1-q)\,F\left(\beta_1^L\right)\right] \cdot P - d}_{E[\pi|\text{explore}]} > \underbrace{F\left(0\right) \cdot P}_{E[\pi|\text{abandon}]} \tag{4.1}$$

[3] For the purposes of this illustrative model, I treat α as fixed. If α were endogenous and costless, the player's optimal α would be infinite, since the exploration upside would then be unlimited and the downside bounded at zero. A natural extension would be to endogenize α and allow exploration costs $d(\cdot)$ or the probability of a successful outcome $q(\cdot)$ to vary with it. Such a model is considerably more difficult to study and beyond the scope of this paper.

These conditions can be rearranged and be written as follows:

$$qF\left(\beta_1^H\right)+(1-q)\,F\left(\beta_1^L\right)-F\left(\beta_0\right)>\frac{d-c}{P} \tag{3.2}$$

$$qF\left(\beta_1^H\right)+(1-q)\,F\left(\beta_1^L\right)-F\left(0\right)>\frac{d}{P} \tag{4.2}$$

In words, the probability gains from exploration over exploitation or abandonment must exceed the difference in cost, normalized by the prize. If the difference in the cost of exploration versus exploitation is small relative to the prize, as it likely is in the data, the choice between them reduces to a question of which choice yields the greater increase in the player's probability of winning.

Effects of Competition

This modeling infrastructure leads directly to the focal propositions, which bring into focus how competition directly affects incentives for exploration.[4] To simplify the presentation, we will assume $d = c$, although the core result (that exploration is incentivized at intermediate levels of competition) also holds when $d > c$, with slightly more involved propositions, provided that d is not so high that exploration will never be preferred to the alternatives (see Appendix A). The first proposition states that when μ_j is high, exploration has greater expected benefits than exploitation, whereas when μ_j is low, the reverse holds. The second proposition states that as μ_j grows large, the benefits of a second design decline to zero. Because effort is costly, players are therefore likely to abandon the contest when competition grows severe.

Proposition 1. *Suppose $q \in \left(\frac{1}{1+\alpha}, \frac{1}{2}\right)$. Then, there exists a μ^* such that for all $\mu_j < \mu^*$,*

$$\underbrace{F\left(\beta_{j0}\right)}_{E[\Pr(\text{Win})|\text{exploit}]} \quad > \quad \underbrace{\left[qF\left(\beta_{j1}^H\right)+(1-q)\,F\left(\beta_{j1}^L\right)\right]}_{E[\Pr(\text{Win})|\text{explore}]}$$

and for all $\mu_j > \mu^*$,

$$\underbrace{\left[qF\left(\beta_{j1}^H\right)+(1-q)\,F\left(\beta_{j1}^L\right)\right]}_{E[\Pr(\text{Win})|\text{explore}]} \quad > \quad \underbrace{F\left(\beta_{j0}\right)}_{E[\Pr(\text{Win})|\text{exploit}]}$$

Proposition 2. *The returns to a player's second design decline to zero as $\mu_j \longrightarrow \infty$.*

Proofs are provided in Appendix A. The necessary condition for competition to motivate exploration is that $q \in \left(\frac{1}{1+\alpha}, \frac{1}{2}\right)$, which holds if and only if original submissions are in expectation higher-quality than tweaks, but successful outcomes are nevertheless improbable (see Appendix A) – in other words, that exploration

[4] The propositions are provided in partial equilibrium (i.e., without strategic interactions) to emphasize the first-order tradeoffs faced by agents in this setting. Strategic interactions, however, would not affect the result: at very small or large values of μ, competitors' best responses will have little influence on the shape of the focal player's success function, and therefore little influence on the difference in returns to exploration versus exploitation. In the middle, there exists a threshold μ^* that divides the real line into regions where exploration or exploitation yields greater benefits.

is not only risky, but also high-return. When this is the case, the first proposition shows that competition can provoke exploration as a strategic response, a result which is similar to the findings of Cabral (2003) and Anderson and Cabral (2007) on choices over risk, but in a structure more closely linked to the empirical setting: intuitively, when the player lags behind, the upside to exploration grows more valuable and the downside less costly. The second proposition shows, however, that large performance differences can also discourage effort, as the returns to effort decline to zero. The proposition is a reminder that participation must be incentivized: in contrast to many bandit models or models of choices over risk in competition (e.g., Cabral 2003), agents in this setting incur costs and may withhold effort.[5]

2 Setting, Data, and Identification

I collect a randomly-drawn sample of 122 logo design contests from a widely-used online platform to study how creative behavior responds to competition.[6] The platform from which the data were collected hosts hundreds of contests each week in several categories of commercial graphic design, including logos, business cards, t-shirts, product packaging, book/magazine covers, website/app mockups, and others. Logo design is the modal design category on this platform and is thus a natural choice for analysis. A firm's choice of logo is also nontrivial, since it is the defining feature of its brand, which can be one of the firm's most valuable assets and is how consumers will recognize and remember the firm for years to come.

In these contests, a firm (the sponsor; typically a small business or non-profit organization) solicits custom designs from freelance designers (players) in exchange for a fixed prize awarded to its favorite entry. The sponsor publishes a project brief which describes its business, its customers, and what it likes and seeks to communicate with its logo; specifies the prize structure; sets a deadline for submissions; and opens the contest to competition. While the contest is active, players can enter (and withdraw) as many designs as they want, at any time they want, and sponsors can provide players with private, real-time feedback on their submissions in the form of 1- to 5-star ratings and written commentary. Players see a gallery of competing designs and the distribution of ratings on these designs, but not the ratings on specific competing designs. Copyright is enforced.[7] At the end of the contest, the sponsor picks the winning design and receives the design files and full rights to their use. The platform then transfers payment to the winner.

For each contest in the sample, I observe the project brief, which includes a project title and description,

[5] Altogether, the model proposes that incentives for exploration are greatest at intermediate levels of competition (see Appendix A). Mathematically, the result is driven by the curvature of the success function, which rises and then flattens with competition. Only at intermediate levels of competition does the function have adequate curvature to make the returns to exploration both larger than those to exploration and large enough to exceed the cost.

[6] The sample consists of all logo design contests with public bidding that began the week of Sept. 3-9, 2013 and every three weeks thereafter through the week of Nov. 5-11, 2013, excluding those with multiple prizes or mid-contest rule changes such as prize increases or deadline extensions. Appendix B describes the sampling procedures in greater detail.

[7] Though players can see competing designs, the site requires that all designs be original and enforces copyright protections. Players have numerous opportunities to report violations if they believe a design to be copied or otherwise misused. Violators are permanently banned from the site. The site also prohibits the use of stock art and has a strict policy on the submission of overused design concepts. These mechanisms appear to be effective at limiting abuses.

the sponsor's industry, and any specific elements that must be included in the logo; the contest's start and end dates; the prize amount; and whether the prize is committed (the sponsor may retain the option of not awarding the prize to any entries if none are to its liking). While multiple prizes are possible, the sample is restricted to contests with a single, winner-take-all prize. I also observe every submitted design, the identity of the designer, his or her history on the platform, the time at which the design was entered, the rating it received (if any), the time at which the rating was given, and whether it won the contest. I also observe when players withdraw designs from the competition, but I assume withdrawn entries remain in contention, as sponsors can request that any withdrawn design be reinstated. Since I do not observe written feedback, I assume the content of written commentary is fully summarized by the rating.[8]

The player identifiers allow me to track players' activity over the course of each contest. I use the precise timing information to reconstruct the state of the contest at the time each design is submitted. For every design, I calculate the number of preceding designs in the contest of each rating. I do so both in terms of the feedback available (i.e., observed) at the time of submission as well as the feedback eventually provided. To account for the lags required to produce a design, I define preceding designs to be those entered at least one hour prior to a given design, and I similarly require that feedback be provided at least one hour prior to the given design's submission to be considered observed at the time it is made.

The dataset also includes the designs themselves. Recall that creativity bears the formal definition of the act of producing ideas that are novel and relevant to the goal at hand (Amabile 1996, Sternberg 2008). To operationalize this definition, I invoke image comparison algorithms commonly used in content-based image retrieval software (similar to Google Image's Search by Image feature) to measure the similarity of each design entered into a contest to preceding designs by the same and other players. I use two mathematically distinct procedures to compute similarity scores for image pairs, one of which is a preferred measure (the "perceptual hash" score) and the other of which is reserved for robustness checks (the "difference hash" score). Appendix B explains how they work. Each algorithm takes a pair of digital images as inputs, summarizes them in terms of a specific, structural feature, and returns a similarity score in the [0,1] interval, with a value of one indicating a perfect match and a zero indicating total dissimilarity. This index effectively measures the absolute correlation of two images' underlying structure, reflecting similarities or differences in the basic shapes, outlines, and other elements that define the image.

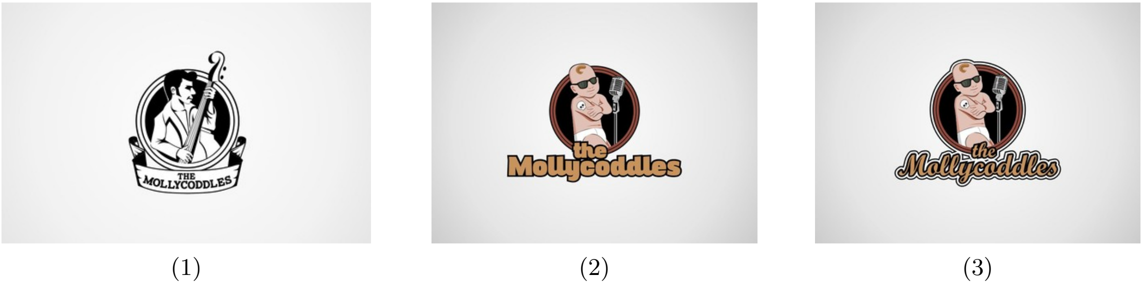
To make this discussion concrete, the inset below demonstrates an example application. The figure shows three designs, entered in the order shown, by the same player in a logo design competition that is similar to those in the sample, although not necessarily from the same platform.[9] The first two logos have some features in common (they both use a circular frame and are presented against a similar backdrop), but they also have some stark differences. The perceptual hash algorithm gives them a similarity score of 0.31, and

[8] One of the threats to identification throughout the empirical analysis is that the estimated effects of ratings may be confounded by unobserved, written feedback: what seems to be a response to a rating could be a reaction to explicit direction provided by the sponsor. This concern is evaluated in detail in Appendix D and discussed later in the paper.

[9] To keep the platform from which the sample was collected anonymous, I omit identifying information.

the difference hash algorithm scores them 0.51. The latter two logos are more alike, and though differences remain, they are now more subtle and mostly limited to the choice of font. The perceptual hash algorithm gives these logos a similarity score of 0.71, and the difference hash scores them 0.89.

Illustration of image comparison algorithms



(1) (2) (3)

Notes: Figure shows three logos entered in order by a single player in a single contest. The perceptual hash algorithm calculates a similarity score of 0.313 for logos (1) and (2) and a score of 0.711 for (2) and (3). The difference hash algorithm calculates similarity scores of 0.508 for (1) and (2) and 0.891 for (2) and (3).

For each design in a contest, I compute its maximal similarity to previous designs in the same contest by the same player. Subtracting this value from one yields an index of originality between 0 and 1, which can be interpreted as an empirical counterpart to the parameter $^1/_\alpha$ in the model. In the empirical analysis, I primarily use measures of similarity to a player's *highest-rated* previous submissions, rather than all of her prior submissions, but since players tend to re-use only their highest-rated work, these two measures are highly correlated in practice ($\rho = 0.9$ under either algorithm).

Creativity can manifest in this setting in other ways. For example, players sometimes create and enter several designs at once, and when doing so they can make each one similar to or distinct from the others. To capture this phenomenon, I define "batches" of proximate designs entered into the same contest by a single player and compute the maximum intra-batch similarity as a measure of creativity in batched work. Two designs are proximate if they are entered within 15 minutes of each other, and a batch is a set of designs in which every design in the set is proximate to another in the same set.[10] Intra-batch similarity is arguably closer to a true measure of experimentation, reflecting players' tendency to try minor variants of the same concept versus multiple concepts over a short period of time.

These measures are not without drawbacks or immune to debate. One drawback is that algorithmic comparisons require substantial dimensionality reduction and thus provide only a coarse comparison between designs based on a select set of features. Concerns on this front are mitigated by the fact that the empirical results throughout the paper are similar in sign, significance, and magnitude under two distinct algorithms. In addition, coarse comparisons will be sufficient for detecting designs that are plainly tweaks to earlier work versus those that are not, which is the margin that matters most for this paper. One may also question how well the algorithms emulate human perception, but the example provided above assuages this concern, as

[10] Note that all batch-level results are similar when defining batches based on 5-, 15-, 60-, or 180-minute intervals.

do other examples in Appendix B, which discusses these issues in detail.

2.1 Characteristics of the Sample

The average contest in the data lasts eight days, offers a $250 prize, and attracts 96 designs from 33 players (Table 1). On average, 64 percent of designs are rated; less than three receive the top rating.

[Table 1 about here]

Among rated designs, and the median and modal rating is three stars (Table 2). Though fewer than four percent of rated designs receive a 5-star rating, over 40 percent of all winning designs are rated five stars, suggesting that these ratings convey substantial information about a design's quality and odds of success.[11] The website also provides formal guidance on the meaning of each star rating, which generates consistency in their interpretation and use across different sponsors and contests.

[Table 2 about here]

Table 3 characterizes the similarity measures. For each design in the sample, we can compute its maximal similarity to previous designs by the same player, the highest-rated previous designs by the same player, and the highest-rated previous designs by that player's competitors (in the same contest). For every design batch, I calculate the maximal similarity of any two designs in that batch. Note that the analysis of intra-batch similarity is restricted to batches that are not missing any image files.

[Table 3 about here]

The designs themselves are available for 96 percent of submissions in the sample. The table shows that new entries are on average more similar to that player's own designs than her competitors' designs, and that designs in the same batch tend to be more similar to each other than to previous designs by even the same player. But these averages mask more important patterns at the extremes. At the upper decile, designs can be very similar to previous work by the same player (≈ 0.75 under the perceptual hash algorithm) or to other designs in the same batch (0.91), but even the designs most similar to competing work are not all that similar (0.27). At the lower end, designs can be original by all of these measures.

Are ratings meaningful? Evidence from the empirical success function

A simple cross-tabulation of ratings suggests that they are meaningful: in roughly two-thirds of contests, the winning design also had the highest rating in that contest (Appendix Table C.3). But the relative value

[11] Another 33 percent of winning designs are rated 4 stars, and 24 percent are unrated.

of each rating can be obtained by estimating an empirical success function, using the win-lose outcomes of each design in a large sample of contests from this platform, which I borrow from Gross (2017).[12] Recall from Section 1 that a design's latent value is a function of its rating and an i.i.d. extreme value error. In the data, there are five possible ratings, such that this latent value can be flexibly specified with fixed effects for each rating and the success function estimated as a simple conditional logit. To formalize, let R_{ijk} denote the rating on design i by player j in contest k, and (in a slight abuse of notation) let $R_{ijk} = \emptyset$ when design ijk is unrated. The value of each design, ν_{ijk}, can be written as follows:

$$\nu_{ijk} = \gamma_{\emptyset}\mathbb{1}(R_{ijk} = \emptyset) + \gamma_1\mathbb{1}(R_{ijk} = 1) + \ldots + \gamma_5\mathbb{1}(R_{ijk} = 5) + \varepsilon_{ijk} \equiv \psi_{ijk} + \varepsilon_{ijk}$$

The details of the estimation are provided in Appendix C, but here we can summarize the results, which will be used in later analysis. The fixed effects are monotonically increasing in the rating and precisely estimated, and only a 5-star design is on average preferred to the outside option. To produce the same increase in a player's estimated win probability as generated by a five-star design, a player would need 12 four-star designs, 137 three-star designs, or nearly 2,000 one-star designs. The magnitudes of these differences imply that for players with a top rating, competitive pressure primarily comes from other top-rated designs. As a measure of fit, the odds-on favorite wins almost half of all contests in the sample. With knowledge of the distribution of their competitors' ratings, which is observable in practice, players can thus invoke a simple heuristic model similar to the one estimated here in their decision-making.

2.2 Empirical Methods and Identification

I exploit variation in the level and timing of ratings to estimate the effects of feedback and competition on players' creative choices. With timestamps on all activity, I can determine exactly what a player knows at each point in time about their own and their competitors' performance and identify the effects of performance differences known to players at the time of design. Identification is thus achieved by harnessing variation in the *information* players possess about the state of the competition.

Concretely, the analysis compares the actions of players who have received the top rating or who know they face top-rated competition aginst those who do not – whether it is because no prior designs will be given the top rating, or because these ratings have simply not yet been administered. The identifying assumption is that there are no omitted factors correlated with observed feedback that also affect choices. This assumption is supported by two pieces of evidence. First, the arrival of ratings is unpredictable, such that the set of ratings observed at any point in time is effectively random: sponsors are erratic, and it is difficult to know exactly when or how often a sponsor will log onto the site to rate new entries, much less any single design.

[12] Estimating the success function requires a larger sample of winners, and thus contests, than are in the primary sample of this paper. As Appendix C shows, the sample in Gross (2017) contains >4,000 contests, is empirically comparable, and includes all of the same variables except for the images themselves – sufficient for the exercise.

More importantly, players' choices are uncorrelated with ratings that were unobserved at the time, including forthcoming ratings and ratings on specific competing submissions.[13]

To establish that feedback provision is unpredictable, I explore the relationship between feedback lags and ratings. In concept, sponsors may be quicker to rate the designs they like the most, to keep these players engaged and improving their work, in which case players might be able to infer their eventual ratings from the time elapsed without any feedback. Empirical assessment of this question (Appendix D) confirms that this is not the case: whether measured in hours or as a percent of the total contest duration, the lag between when a design is entered and rated is unrelated to its rating. The probability that a rated design was rated before versus after the contest ends is similarly unrelated to the rating granted.

Evidence that choices are uncorrelated with unobserved or not-yet-observed ratings is presented in Section 3. Following the discussion of the focal results, I show that the relationship between the similarity measures and forthcoming ratings is indistinguishable from zero. In unreported tests, I also examine players' tendency to imitate highly-rated competing designs and find no such patterns, likely because they simply do not know which designs are highly rated (and thus which ones to imitate). These collective results suggest that players respond only to ratings observed at the time of design.

Finally, a distinct threat to identification arises if ratings are accompanied by written feedback, and these comments provide explicit instruction that generates the patterns found in this paper. Appendix D evaluates this possibility in detail, using a newly-drawn sample of contests in which written comments were made visible to the public, seemingly by error. Within this sample, sponsors provided written comments to fewer than 8 percent of submissions, though the frequency is significantly higher for highly-rated submissions than for poorly-rated submissions. These comments take a range of flavors, with many echoing the rating given, but some make a more explicit request or suggestion of content changes. Because the latter present the risk of confounding the results of this paper, I hired individuals to read every comment in this sample and determine whether the sponsor suggested specific changes. Using this measure, I find that instructive comments are (i) rare, (ii) not disproportionately provided to 5-star designs relative to 3- or 4-star designs, and most importantly (iii) not related to the presence of high-rated competition, such that they cannot be responsible for differences in behavior that will be found in the analysis below.

Interpretation: Feedback versus Market Structure

Although the phenomenological focus of the paper is on the effects of market structure on creativity, the identifying variation is generated by feedback, raising the question of whether the empirical results should be attributed to "feedback" or to "market structure" – though the distinction is blurred by the fact that in this setting, feedback *is* information about market structure. To orient this question, it is helpful to observe

[13] Though this setting may seem like a natural opportunity for a controlled experiment, the variation of interest is in the 5-star ratings, which are sufficiently rare that a controlled intervention would require either unrealistic manipulation or an infeasibly large sample. I therefore exploit naturally-occurring variation for this study.

that feedback can affect choices through two channels in this setting. The rating on a given design will (i) help that player understand whether it is good or bad, and thereby project how her next design might be rated if she enters something similar to it versus different. But that rating, together with knowledge of the ratings given to competitors, will also (ii) inform her whether her probability of winning is high or low – and if the player's objective is to win, this probability is fundamentally what will be driving behavior in these contests, as in the model in Section 1. In the following sections, I therefore refer to the effect of a player's own ratings as an effect of *feedback* or (for example) of a *high rating*, and to the effect of competitors' ratings as an effect of *high-rated competition*, with the primary focus being the latter.

3 Competition and Creativity

Figure 1 provides a first-cut, visual preview of this section's results. The figure shows the distribution of designs' maximal similarity to previous submissions by the same player, conditioning on that player's best rating and whether top-rated competition was present at the time. Recall that low similarity scores indicate that the submission is substantively original, while high scores indicate that it is a variant on a prior entry. Submissions from players with high ratings are significantly more similar to their prior work, with the effects largest for those with a 5-star rating (bottom-right panel). However, these very same players are more likely to enter original designs when facing top-rated competition than in its absence.

[Figure 1 about here]

The richness of the field setting makes it possible to evaluate whether competition affects high-performing players' tendency to enter more original work in a variety of ways. The estimating equation in this part of the paper is as follows, with variants estimated throughout this section:

$$\begin{aligned} Similarity_{ijk} \quad = \quad & \alpha + \textstyle\sum_{r=2}^{5} \beta_r \cdot \mathbb{1}(\bar{R}_{ijk} = r) \\ & \qquad + \beta_{5c} \cdot \mathbb{1}(\bar{R}_{ijk} = 5) \cdot \mathbb{1}(\bar{R}_{\text{-}ijk} = 5) \\ & \qquad + \beta_{5p} \cdot \mathbb{1}(\bar{R}_{ijk} = 5) \cdot P_k \\ & \quad + \beta_c \cdot \mathbb{1}(\bar{R}_{-ijk} = 5) + \beta_p \cdot P_k \\ & \quad + T_{ijk}\lambda + X_{ijk}\theta + \zeta_k + \varphi_j + \varepsilon_{ijk} \end{aligned}$$

where i indexes designs by player j in contest k. The variables are as follows: $Similarity_{ijk}$ is the maximal similarity of design ijk to the highest-rated preceding designs by player j in contest k; $\bar{R}_{ijk}$ is the highest rating received by player j in contest k prior to design ijk; $\bar{R}_{\text{-}ijk}$ is the highest rating received by player j's competitors prior to design ijk; P_k is the prize in contest k, in units of \$100s; T_{ijk} is fraction of the contest elapsed when design ijk is entered; X_{ijk} consists of other design-level controls, including counts of previous designs by the same player and by competing players, and the number of days remaining in the contest; and ζ_k and φ_j are contest and player fixed effects. Specifications without contest fixed effects include the prize P_k as a standalone explanatory variable to identify the interaction. Standard errors throughout are clustered by player to account for any within-player correlation in the error term, though the results are robust to (and more conservative than) clustering by contest-player or by contest.

It may be helpful to provide a roadmap to this part of the analysis in advance. In the first set of regressions, I estimate the specification above. The second set estimates a specification in first differences, replacing the dependent variable with the change in similarity to previously-rated designs and independent variables with indicators for changes in highest ratings. The third set studies within-batch similarity. The fourth set tests the identifying assumption that players do not act on forthcoming ratings. The fifth set explores whether the effects of competition are general to players with lower ratings when no 5-star ratings have been granted in a contest. The final set evaluates the similarity of players' initial submissions in each contest to designs from other players and contests, as these submissions are mechanically excluded from the preceding analysis. Additional robustness checks are provided in the appendix.

3.1 Similarity of new designs to a player's previous designs

I begin by examining players' tendency to enter novel versus derivative designs. Table 4 provides estimates from regressions of the maximal similarity of each design to the highest-rated preceding designs by the same player on indicators for the highest rating received. All specifications interact the indicator for the top rating with (i) the prize (in $100s) and (ii) a variable indicating the presence of top-rated competition, and control for the fraction of the contest elapsed. Column (1) presents a baseline with no fixed effects or other controls. Columns (2) and (3) add fixed effects for contests and players, respectively, and Column (4) includes both. Column (5) additionally controls for the number of days remaining in the contest and the number of prior submissions by the same player as well as competing players.

[Table 4 about here]

Similar patterns are observed across all specifications. In the regression with both fixed effects and controls (Column 5), we see that players with the top rating enter designs that are 0.36 points, or over one full standard deviation, more similar to their previous work than players who have only low ratings or no ratings. Roughly half of this effect is reversed by the presence of top-rated competition, with this counteracting effect significant at the one percent level. When a player's highest rating is four stars, her new designs are on average around 0.1 points more similar to previous work. This effect further attenuates as the best observed rating declines until it is indistinguishable from zero at a best rating of two stars, with all such differences statistically significant. High-rated competition is not observed to have an effect on similarity for these lower performers, who are already unlikely to reuse their low-rated submissions.

The latter regressions in Table 4 use contest and player fixed effects to control for other factors that are either common to all players within a contest or across all contests for a given player, but they do not control for factors that are constant for a given player within specific contests, as doing so leaves too little variation to identify the focal effects. Such factors could nevertheless be confounding, such as if players who continue participating in different competitive conditions are systematically more or less likely to enter similar designs in that contest. The estimates in the previous tables additionally mask potential heterogeneity in players' reactions to competitive conditions over the course of a contest.

Table 5 addresses these concerns with a model in first differences. The dependent variable here is the *change* in designs' similarity to the player's best previously-rated work. This variable can take values in [-1,1], where a value of 0 indicates that the given design is as similar to the player's best preceding design as was the last one she entered; a value of 1 indicates that the player transitioned fully from innovating to recycling; and a value of -1, the converse. The independent variables are changes in indicators for the highest rating the player has received, interacting the indicator for the top rating with the prize and the presence of top-rated competition. I estimate this model with the same configurations of contest fixed effects, player fixed effects, and controls to account for other potential reasons why players' propensity for similarity changes over time, though the results are not statistically different across these specifications.

[Table 5 about here]

The results provide even stronger evidence of how competition affects creative choices and are statistically and quantitatively similar across specifications. A player who receives her first 5-star rating will typically then enter a near replica: the similarity increases by 0.9 points, or over *three* standard deviations, relative to players with low ratings. Top-rated competition again reverses roughly half of this effect, with the difference significant at the one percent level. Given their magnitudes, these effects will be plainly visible to the naked eye (the inset in Section 2 gives an example of what they would look like in practice). The effects of a new best rating of four, three, or two stars attenuate monotonically, similar to earlier results, and high-rated competition is not seen to have an effect on low performers.[14]

Table 6 tests the effects of competition in a different place: within batches of two or more designs entered by a given player, in a given contest, at once or in rapid succession. Recall that when entering multiple designs at once, players can make them similar to or different from each other.[15] Observations here are submission

[14] Interestingly, these regressions also find that new recipients of the top rating can also be induced to try new designs with larger prizes. The theory suggests a possible explanation: large prizes moderate the influence of costs in players' decision-making. If original designs are more costly (take more time or effort) than tweaks, they may be more worth doing when the prize is large. This is particularly the case for players with highly-rated work in the contest.

[15] Note that submitting two similar designs in succession can be a low-risk or a high-risk move, depending on the circumstances: if they are both tweaks on a third design, it is a low-variance play; if the first design is novel and the second is a tweak of it, it is a high-variance, eggs-in-one-basket move with compounded risk. Empirically, we see that when two designs in a batch are similar to each other, they are also similar to a third design by the same player which preceded the batch, supporting the interpretation of high intra-batch similarity as risk reduction.

batches, and the dependent variable is the maximal similarity of any two designs in the batch, estimated as a function of the ratings observed at the time of submission.

[Table 6 about here]

The results indicate that maximal within-batch similarity declines 0.3 points, or one standard deviation, for players with the top rating who face top-rated competition, relative to those who do not. The effect is insensitive to inclusion of controls (Columns 2 and 4) or weighting batches by their size (Columns 3 and 4), albeit significant only at the 10 percent level. High-rated players with competition are thus more likely to produce original designs not only across batches but also within them.

The appendix provides robustness checks and supplementary analysis. To confirm that these patterns are not an artifact of the perceptual hash algorithm, Appendix E re-estimates the regressions in the preceding tables using the difference hash algorithm to calculate similarity scores. The results are both statistically and quantitatively similar. In Appendix F, I split out the effects of competition by the number of top-rated competing designs, finding no statistical differences between the effects of one versus more than one: all of the effects of competition are achieved by one high-quality competitor.

This latter result is especially important for ruling out an information-based story. In particular, the presence of other 5-star ratings might indicate that the sponsor has diverse preferences, and that unique designs have higher likelihood of being well-received than one might otherwise believe. If this were the case, then similarity should continue to decline as 5-star competitors are revealed. That this is not the case suggests that the effect is in fact the result of variation in incentives from competition.

In unreported tests, I also look for effects of 5-star competition on players with only 4-star designs, and find attenuated effects that are negative but not significantly different from zero. I also explore the effects of prize commitment, since the sponsor's outside option of not awarding the prize is itself a competing alternative. The effect of prize commitment is not statistically different from zero. I similarly test for effects of four-star competition on players with five-star designs, finding none. These results reinforce the earlier evidence that competition effectively comes from other designs with the top rating.[16]

3.2 Placebo test: Similarity to a player's not-yet-rated designs

The identifying assumptions require that players are not acting on information that correlates with feedback but is unobserved in the data. As a simple validation exercise, the regressions in Table 7 perform a placebo test of whether similarity is related to impending feedback. If an omitted determinant of creative choices is correlated with ratings and biasing the results, then it would appear as if similarity responds to forthcoming ratings. If the identifying assumptions hold, we should see only zeros.

[16] In additional unreported results, I also re-estimate Table 4 for players who entered the contest when the highest competing rating was 4-stars or higher versus 3-stars or lower, to see whether selection into the contest on the intensity of competition might explain the results, and find similar effects for both groups.

[Table 7 about here]

The specification in Column (1) regresses a design's maximal similarity to the player's best designs that have not yet been *but will eventually be* rated on indicators for the ratings they later receive. I find no evidence that designs' similarity is related to forthcoming ratings. Because a given design's similarity to an earlier, unrated design can be incidental if both are tweaks on a third design, Column (2) adds controls for similarity to the best already-rated design. Column (3) allows these controls to vary with the rating received. As a final check, I isolate the similarity to the unrated design that cannot be explained by similarity to the third design in the form of a residual, and in Column (4) I regress these residuals on the same independent variables. In all cases, I find no evidence that players systematically tweak designs with higher forthcoming ratings. Choices are only correlated with ratings observed in advance.

3.3 Extensions: 4-on-4 Competition

The model suggests that similar dynamics should arise for players with 4-star ratings facing 4-star competition when no higher ratings are granted, since only relative performance matters – though this result may only arise towards the end of a contest, when the absence of higher ratings approaches finality. Table 8 tests this prediction, regressing designs' similarity to the player's prior entries on indicators for their highest rating, restricting to submissions made before any 5-star designs were in play.

[Table 8 about here]

Column (1) includes all designs in the sample that meet this condition. Column (2) restricts to submissions in the second half of a contest; Column (3), to the final quarter. The table shows similar patterns for 4-on-4 competition that strengthen over the course of a contest, though the sample in the most restrictive condition is sufficiently small that standard errors cannot rule out no effect.[17] For comparison, 4-star competition does not have a comparable effect when 5-star competition is already present, nor on players who themselves have a 5-star rating. In addition to extending the main findings, these results thus reinforce the evidence that the observed behavior is a response to incentives (driven by relative performance differences), rather than information about sponsors' preferences (from competitors' absolute ratings).

3.4 Extensions: Initial Submissions

One limitation of these results is that they only examine players' second and later submissions, since the similarity measure requires at least one previous design to compare against. Given that most players enter

[17] In the data, contests rarely reach the final quarter with no players having received a 5-star rating and only one player having received a 4-star rating. As a result, there is only a small sample off which the focal coefficients (on the 4-star rating and its interaction with 4-star competition) in Column (3) can be estimated, and the estimates are imprecise.

multiple designs, this nevertheless comprises the majority of submissions – but a full third of designs in the data are first submissions. Can anything be said about these designs?

Measurement is complicated by the fact that there is no obvious precedent to compare against – in a sense, all first submissions are original to a contest, provided they do not imitate competing entries, which is both prohibited by the platform and rare in practice. But to look for patterns, I consider four approaches: (i) compare initial submissions to the highest-rated prior designs in the same contest, (ii) compare them to all prior designs in the same contest, (iii) compare them to all designs by the same player in *other* contests in the data, and (iv) compare them to all designs by *any* player in other contests. The latter two comparisons are intended as measures of "overall" originality, as measured against a player's own portfolio or the entire platform. In all cases, I regress the similarity measure on indicators for the highest competing rating in the contest at the time of entry, plus the fixed effects and controls from previous specifications, and all regressions are conditional on at least one rating having already been granted in the contest. These variants are shown in Columns (1) through (4) of Table 9 below.

[Table 9 about here]

The table shows few significant patterns, except that initial submissions may slightly deviate away from high-rated competitors (Column 1) – though it is difficult to put too much weight on this result, given that players cannot see which competing design is highest-rated, and because it does not carry over to comparisons against all prior designs in a contest (Column 2). The effects documented in previous tables thus appear limited to similarities and differences within a player's line of work in a given contest.

4 Effects on Participation

The analysis thus far conditions on continued participation: the unit of observation is a submission, and the outcome is its similarity to previous submissions. However, players can also stop making submissions if they perceive the returns to effort to be low. This outside option is present in many real-world competitive settings, and it distinguishes the setting of this paper from much of the existing literature on creativity, innovation, and high-powered incentives, where agents are effectively locked in to participating.

To incorporate the outside option into the empirics, I discretize outcomes and model each submission as a choice between three options: (i) entering a tweak and remaining active ("tweak"), (ii) entering an original design and remaining active ("original"), or (iii) entering any design and refraining from further submissions ("abandon"). Although the precise moment that a player decides to stop investing effort is not observable, we can use inactivity as a proxy. The unit of observation thus remains an individual submission, but I now categorize designs as original or as a tweak on the basis of discrete cutoffs for similarity scores, and I identify designs which are its creator's final submission into each contest.[18]

The multinomial choice framework is necessary because the tradeoffs between three unordered options (tweak, original, or abandon) cannot be evaluated in a linear model. For this exercise, I classify a design as a tweak if its similarity to any earlier design by the same player is 0.7 or higher and original if its maximal similarity to previous designs by that player is 0.3 or lower.[19] Designs with intermediate similarity scores are omitted from the exercise, as the player's intentions are ambiguous in the intermediate range. Each action a in this choice set is assumed to have latent utility u^a_{ijk} for submission i by player j in contest k. Much like the previous specifications, I model this latent utility as a function of the player's own ratings, interacting the indicator for the top rating with indicators for 1, 2, and 3+ top-rated competitors, the fraction of the contest transpired, the number of days remaining, and a logit error term:

$$\begin{aligned} u^a_{ijk} \quad = \quad & \beta^a_0 + \textstyle\sum_{r=1}^{5} \beta^a_r \cdot \mathbb{1}(\bar{R}_{ijk} = r) \\ & \quad + \beta^a_{5,1} \cdot \mathbb{1}(\bar{R}_{ijk} = 5)\mathbb{1}(N_{\text{-}ijk} = 1) \\ & \quad + \beta^a_{5,2} \cdot \mathbb{1}(\bar{R}_{ijk} = 5)\mathbb{1}(N_{\text{-}ijk} = 2) \\ & \quad + \beta^a_{5,3} \cdot \mathbb{1}(\bar{R}_{ijk} = 5)\mathbb{1}(N_{\text{-}ijk} \geq 3) \\ & + \gamma^a_1 \cdot \mathbb{1}(N_{\text{-}ijk} = 1) + \gamma^a_2 \cdot \mathbb{1}(N_{\text{-}ijk} = 2) + \gamma^a_3 \cdot \mathbb{1}(N_{\text{-}ijk} \geq 3) \\ & + T_{ijk}\lambda^a + X_{ijk}\theta^a + \varepsilon^a_{ijk}, \quad \varepsilon^a_{ijk} \sim \text{i.i.d. Type-I E.V.} \end{aligned}$$

where $N_{\text{-}ijk}$ is the number of top-rated competing designs at the time submission ijk is made, X_{ijk} controls solely for the number of days remaining in the contest, and the other variables are defined as before. Controlling for the fraction of the contest transpired and number of days remaining is especially important here, as abandonment is mechanically more likely to be observed later in a contest.

I estimate the parameters by maximum likelihood using observed behavior. I then use the results to compute the probability that a player with the top rating takes each of the three actions near the end of a contest, and to evaluate how these probabilities vary as the number of top-rated competitors increases from zero to three or more. These probabilities are shown in Figure 2. Panel A plots the probability that the player tweaks; Panel B, that she enters an original design; and Panel C, that she does either but then quits. The bars around each point provide the associated 95 percent confidence intervals.

[Figure 2 about here]

[18] Note that this measure cannot distinguish a player who stops competing immediately after their final submission from one who waits for more information but later stops competing without additional entries. Because the end result is the same, the distinction is not critical for the purposes of this paper, as both behaviors will be influenced by information available at the time of the final submission. Anecdotally, according to some designers on this platform, it is often the case that players will enter their final design knowing it is their final design and not look back.

[19] Because the distribution of similarity scores is continuous in the data, there is not an obvious cutoff for defining tweaks and original designs. The results below are robust to alternatives such as 0.6/0.4 or 0.8/0.2.

The probability that a high-performer tweaks and remains active (Panel A) peaks at 46 percent when there are no 5-star competitors and is significantly lower with non-zero competition, with all differences significant at the one percent level. The probability that the player produces an original design (Panel B) peaks at 52 percent with one 5-star competitor and is significantly lower with zero, two, or three 5-star competitors (differences against zero and three significant at the one percent level; difference against two significant at the ten percent level). Panel C shows that the probability of abandonment increases monotonically in the level of competition, approaching 70 percent with three or more competitors.

These patterns can also be demonstrated by other means. A simple alternative is to model the same discrete choice as a function of a player's contemporaneous probability of winning (which can be computed using the conditional logit estimates described above) rather than directly as a function of ratings. Figure 3 provides estimated choice probabilities under this specification.

[Figure 3 about here]

In Figure 3, the probability that a player tweaks (Panel A) is maximized at 70 percent when she is a strong favorite and declines monotonically to zero with her odds of winning. The probability that the player produces an original design (Panel B) follows a distinct and highly significant inverted-U pattern, peaking at approximately a one-half odds of winning. Finally, the probability that she abandons (Panel C) increases from zero to around 80 percent as her odds of winning fall to zero.

Table 10 shows the underlying source of these patterns, vis-à-vis the estimated mean utility of each action at different win probabilities, with abandonment normalized to zero. Standard errors are shown in parentheses, and the table marks the highest-utility option in each condition. As players' win probability declines: tweaks, original designs, and abandonment (in order) provide the greatest mean utility. To the extent that original designs are uncertain, as the theoretical framework in Section 1 proposes, the evidence also suggests that players increasingly prefer the risky choice over the safe choice as they fall further behind – an intuitive result, albeit one which is distinct from the findings of Genakos and Pagliero (2012). More flexible specifications, including higher-order polynomials in win probability, yield similar patterns.

[Table 10 about here]

The evidence above is an important reminder that players may stop submitting designs when competition grows severe. Taken together, the evidence reveals that incentives for creativity are greatest with balanced competition: too little, and high-performers lack incentive to develop new ideas; too much, and agents stop investing effort altogether. In the data, it appears that creative effort is most attractive to high-rated players when faced off against exactly one high-rated competitor.

5 Evaluating the Returns to Creativity

Why do these players respond to competition by entering more original work? In conversations with creative professionals, including the panelists hired for the exercise below, several claimed that competition requires them to "be bold" or "bring the 'wow' factor," and that it induces them to take on more creative risk. The key assumption of this interpretation is that creativity is both riskier and higher reward than incremental changes – but whether or not this is true is fundamentally an empirical question.

A natural approach to answering this question might be to look at the distribution of sponsors' ratings on original designs versus tweaks, using the same definitions as before, conditioning on the rating of the tweaked design (for tweaks) or the player's highest prior rating (for originals). When we do so, we find that original designs after a low rating are on average higher-rated than tweaks to designs with low ratings, but original designs after a high rating are on average *lower-rated* than tweaks of top-rated designs, raising the question of why a player would deviate from her top-rated work.

The problem with this approach is that ratings are censored: it is impossible to observe improvements above the 5-star rating. With this top-code, original designs will necessarily appear to underperform tweaks of 5-star designs: the sponsor's rating can only go down. The data are thus inadequate for the exercise. To get around the top-code, I hired a panel of five professional graphic designers to independently assess all 316 designs in my sample that were rated five stars by contest sponsors, and I use the panelists' ratings to evaluate whether creativity is in fact a high-risk, high-return activity.

Results from a Panel of Professional Designers

For this exercise, I hired five professional graphic designers at their regular rates to evaluate each design on an extended scale. These ratings were collected though a custom web-based application in which designs were presented in random order and panelists were limited to 100 ratings per day. With each design, the panelist was provided the project title and client industry (excerpted from the source data) and instructed to rate the "quality and appropriateness" of the given logo on a scale of 1 to 10.[20]

To account for differences in the panelists' austerity, I first demean their ratings, in essence removing rater fixed effects. For each design, I then compute summary statistics of the panelists' ratings (mean, median, maximum, and standard deviation). As an alternative approach to aggregating panelists' ratings, I also calculate each design's score along the first component from a principal component analysis. Collectively, these summary statistics characterize a distribution of opinion on a given design.

[20] Appendix G provides more detail on the survey procedure and shows histograms of ratings from each panelist. One panelist was particularly critical with her ratings and frequently ran up against the lower bound. This pattern was evident after the first day of the survey, and the decision was made at that time to exclude this panelist from subsequent analysis, although the results are robust to including ratings from this panelist above the lower bound.

I then identify designs as being tweaks or originals using the definitions above and compare the level and heterogeneity of panelists' ratings on designs of each type. Table 11 provides the results. Designs classified as tweaks are typically rated below-average, while those classified as original are typically above-average. These patterns manifest for the PCA composite, mean, and median panelist ratings; the difference in all three cases is on the order of around half of a standard deviation and is significant at the one percent level. The maximum rating that a design receives from any panelist is also greater for originals, with the difference significant at the one percent level. Yet so is the level of disagreement: the standard deviation across panelists on a given design is significantly greater for original designs than for tweaks. The evidence thus reinforces a possible link between creativity and risk-taking previously suggested by research in other fields, such as social psychology and neuroscience (e.g., Dewett 2006, who finds that a willingness to take risks is positively associated with employees' creativity in the workplace, and Limb and Braun 2008, who show with fMRI data that jazz pianists' prefrontal cortex – the part of the brain responsible for planning, decision-making, and self-regulation – deactivates during improvisation).

[Table 11 about here]

6 Implications and Conclusion

Within this sample of commercial logo design competitions, I thus find that high-powered incentives have nuanced, multifaceted effects on individuals' creative output: some competition is needed to motivate high performers to develop original, untested ideas over tweaking their earlier work, but heavy competition drives them to stop investing altogether. When the two effects are considered in tandem, the evidence indicates that the likelihood that an agent produces original work is greatest with one competitor of similar ability. The results can be rationalized by a model in which creativity is inherently risky, and in which creative effort involves a choice over risk. As such, the paper ties together literatures in the social psychology of creativity and the economics of tournament competition, and provides new evidence on how competition shapes the intensity and direction of individuals' creative production.

The results have direct implications for the use of incentives as a tool for promoting creativity. The foremost lesson is that competition can motivate creativity in professional settings, provided it is balanced. In designing contracts for creative workers, managers ought thus consider incentives for high-quality work relative to that of peers or colleagues, in addition to the more traditional strategy of establishing a work environment with intrinsic motivators such as freedom, flexibility, and challenge. Note that the reward need not be pecuniary: the same intuition applies when workers value recognition or status.

In practice, this 'Goldilocks' level of competition may be difficult to achieve, let alone determine, and finding it would likely require experimentation with the mechanism itself by a principal in another setting. In this paper, the presence of one high-quality competitor was found to be sufficient to induce another high-quality player to produce original designs. A natural conjecture for other settings may be that a few (perhaps even one) competitor of similar ability is enough to elicit creativity, while the presence of many such competitors would be more harmful than helpful for motivating creative output – although the precise thresholds may also depend on other features of the setting, such as the prize distribution.

The results are also relevant to innovation procurement practices, particularly as governments, private foundations, and firms increasingly contract for R&D through prizes and institutionalize prize competition.[21] Yet the applications are more general than R&D prizes alone: as earlier discussion explains, the mechanism in this paper is fundamentally an RFP, a standard contracting device used by firms and government agencies to solicit designs or prototypes of new products, systems, and technologies, with a prize or production contract awarded to the preferred submission. These competitions often take place over multiple rounds, with performance scored between, much like the contests studied here.

Caution is nonetheless warranted in drawing external inference to other procurement settings, as the product being procured in this paper (a logo) is relatively simple, and proposals are heavily tailored to each client. Another potential challenge to external validity is the absence of objective evaluation criteria: the ratings and winner selection are inherently at the sponsor's subjective discretion. Yet in many RFPs, evaluation criteria similarly leave room for subjective judgments or are otherwise opaque to participants. More importantly, the defining feature of the R&D problem is not the ambiguity or clarity of the evaluation criteria, but rather the uncertainty around how any given product design will perform until it is tested and its performance is revealed. This uncertainty is present in all competitive R&D settings.

The final contribution is more methodological in its nature: this paper introduces new tools for measuring innovation in terms of its content. Whereas most recent attempts at content-based analysis of innovation have focused on textual analysis of patents, this paper demonstrates that even unpatentable ideas can quantified, and it exploits a data-rich setting to study how ideas are developed and refined in response to competition. Many other questions about individual creativity and the process of innovation remain open, and this paper provides an example of how this agenda can be pursued.

[21] For example, the U.S. federal government now operates a platform (Challenge.gov) where agencies can seek solutions to both technical and non-technical problems from the public, with hundreds of active competitions and prizes ranging from status only (non-pecuniary) to tens of million dollars. Similar platforms (e.g., Innocentive) are available to organizations outside of the public sector. See Williams (2012) for a review of the literature on R&D prizes.

References

Aghion, Philippe, Nick Bloom, Richard Blundell, Rachel Griffith, and Peter Howitt, "Competition and Innovation: An Inverted-U Relationship," *Quarterly Journal of Economics*, 2005, *120* (2), 701–728.

_ , Stefan Bechtold, Lea Cassar, and Holger Herz, "The Causal Effects of Competition on Innovation: Experimental Evidence," 2014. NBER Working Paper 19987.

Amabile, Teresa M., *Creativity in Context*, Boulder: Westview Press, 1996.

_ and Mukti Khaire, "Creativity and the Role of the Leader," *Harvard Business Review*, 2008, *October.*

_ and _ , "Creativity and the Role of the Leader," *Harvard Business Review*, 2012, *Published Online.*

Anderson, Axel and Luis Cabral, "Go for broke or play it safe? Dynamic competition with choice of variance," *RAND Journal of Economics*, 2007, *38* (3), 593–609.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar, "Large Stakes and Big Mistakes," *Review of Economic Studies*, 2009, *76*, 451–469.

Baik, Kyung Hwan, "Effort Levels in Contests with Two Asymmetric Players," *Southern Economic Journal*, 1994, pp. 367–378.

Baye, Michael R. and Heidrun C. Hoppe, "The Strategic Equivalence of Rent-seeking, Innovation, and Patent-race Games," *Games and Economic Behavior*, 2003, *44* (2), 217–226.

Boudreau, Kevin J. and Karim R. Lakhani, "'Open' Disclosure of Innovations, Incentives and Follow-on Reuse: Theory on Processes of Cumulative Innovation and a Field Experiment in Computational Biology," *Research Policy*, 2015, *44* (1).

_ , _ , and Michael Menietti, "Performance Responses To Competition Across Skill-Levels In Rank Order Tournaments: Field Evidence and Implications For Tournament Design," *RAND Journal of Economics*, 2016, *47*, 140–165.

_ , Nicola Lacetera, and Karim R. Lakhani, "Incentives and Problem Uncertainty in Innovation Contests: An empirical analysis," *Management Science*, 2011, *57* (5), 843–863.

Bradler, Christiane, Susanne Neckermann, and Arne Jonas Warnke, "Incentivizing Creativity: a Large-Scale Experiment with Tournaments and Gifts," 2018. Forthcoming at the Journal of Labor Economics.

Brown, Jennifer, "Quitters Never Win: The (adverse) incentive effects of competing with superstars," *Journal of Political Economy*, 2011, *119* (5), 982–1013.

Brown, Keith C., W. V. Harlow, and Laura T. Starks, "Of Tournaments and Temptations: An Analysis of Managerial Incentives in the Mutual Fund Industry," *Journal of Finance*, 1996, *51* (1), 85–110.

Cabral, Luis, "R&D Competition When Firms Choose Variance," *Journal of Economics & Management Strategy*, 2003, *12* (1), 139–150.

Charness, Gary and Daniela Grieco, "Creativity and Financial Incentives," 2018. Forthcoming at the Journal of the European Economic Association.

Che, Yeon-Koo and Ian Gale, "Optimal Design of Research Contests," *American Economic Review*, 2003, *93* (3), 646–671.

Chevalier, Judy and Glenn Ellison, "Risk Taking by Mutual Funds as a Response to Incentives," *Journal of Political Economy*, 1997, *105* (6), 1167–1200.

Cohen, Wesley M., "Fifty Years of Empirical Studies of Innovative Activity and Performance," *Handbook of the Economics of Innovation*, 2010, *1*, 129–213.

Dewett, Todd, "Exploring the Role of Risk in Employee Creativity," *Journal of Creative Behavior*, 2006, *40* (1), 27–45.

Ederer, Florian, "Feedback and Motivation in Dynamic Tournaments," *Journal of Economics & Management Strategy*, 2010, *19* (3), 733–769.

_ **and Gustavo Manso**, "Is Pay-for-Performance Detrimental to Innovation?," *Management Science*, 2013, *59* (7), 1496–1513.

Eisenberger, Robert and Judy Cameron, "Detrimental Effects of Reward: Reality or Myth?," *American Psychologist*, 1996, *51* (11), 1153–1166.

_ **and** _ , "Reward, Intrinsic Interest, and Creativity: New Findings," *American Psychologist*, 1998, *June*, 676–679.

Erat, Sanjiv and Uri Gneezy, "Incentives for creativity," *Experimental Economics*, 2016, *19*, 269–280.

Florida, Richard and Jim Goodnight, "Managing for Creativity," *Harvard Business Review*, 2005, *July-August.*

Fudenberg, Drew, Richard Gilbert, Joseph Stiglitz, and Jean Tirole, "Preemption, Leapfrogging, and Competition in Patent Races," *European Economic Review*, 1983, *22* (1), 3–31.

Fullerton, Richard L. and R. Preston McAfee, "Auctioning Entry into Tournaments," *Journal of Political Economy*, 1999, *107* (3), 573–605.

Genakos, Christos and Mario Pagliero, "Interim Rank, Risk Taking, and Performance in Dynamic Tournaments," *Journal of Political Economy*, 2012, *120* (4), 782–813.

Gilbert, Richard, "Looking for Mr. Schumpeter: Where are we in the competition-innovation debate?," in Adam B. Jaffe, Josh Lerner, and Scott Stern, eds., *Innovation Policy and the Economy, Volume 6*, Cambridge: The MIT Press, 2006.

Gross, Daniel P., "Performance Feedback in Competitive Product Development," *RAND Journal of Economics*, 2017, *48* (2), 438–466.

Hennessey, Beth A. and Teresa M. Amabile, "Creativity," *Annual Review of Psychology*, 2010, *61*, 569–598.

_ **and Theresa M. Amabile**, "Reward, Intrinsic Motivation, and Creativity," *American Psychologist*, 1998, *June*, 674–675.

Limb, Charles J. and Allen R. Braun, "Neural Substrates of Spontaneous Musical Performance: An fMRI Study of Jazz Improvisation," *PLoS One*, 2008, *3* (2), e1679.

Manso, Gustavo, "Motivating Innovation," *The Journal of Finance*, 2011, *66* (5), 1823–1860.

Schumpeter, Joseph A., *Capitalism, Socialism, and Democracy*, New York: Harper, 1942.

Shalley, Christina E. and Greg R. Oldham, "Competition and Creative Performance: Effects of Competitor Presence and Visibility," *Creativity Research Journal*, 1997, *10* (4), 337–345.

_ **, Jing Zhou, and Greg R. Oldham**, "The Effects of Personal and Contextual Characteristics on Creativity: Where Should We Go from Here?," *Journal of Management*, 2004, *30* (6), 933–958.

Sternberg, Robert J., *Cognitive Psychology*, 5th ed., Belmont: Wadsworth, 2008.

Taylor, Curtis R., "Digging for Golden Carrots: An analysis of research tournaments," *The American Economic Review*, 1995, pp. 872–890.

Terwiesch, Christian and Yi Xu, "Innovation Contests, Open Innovation, and Multiagent Problem Solving," *Management Science*, 2008, *54* (9), 1529–1543.

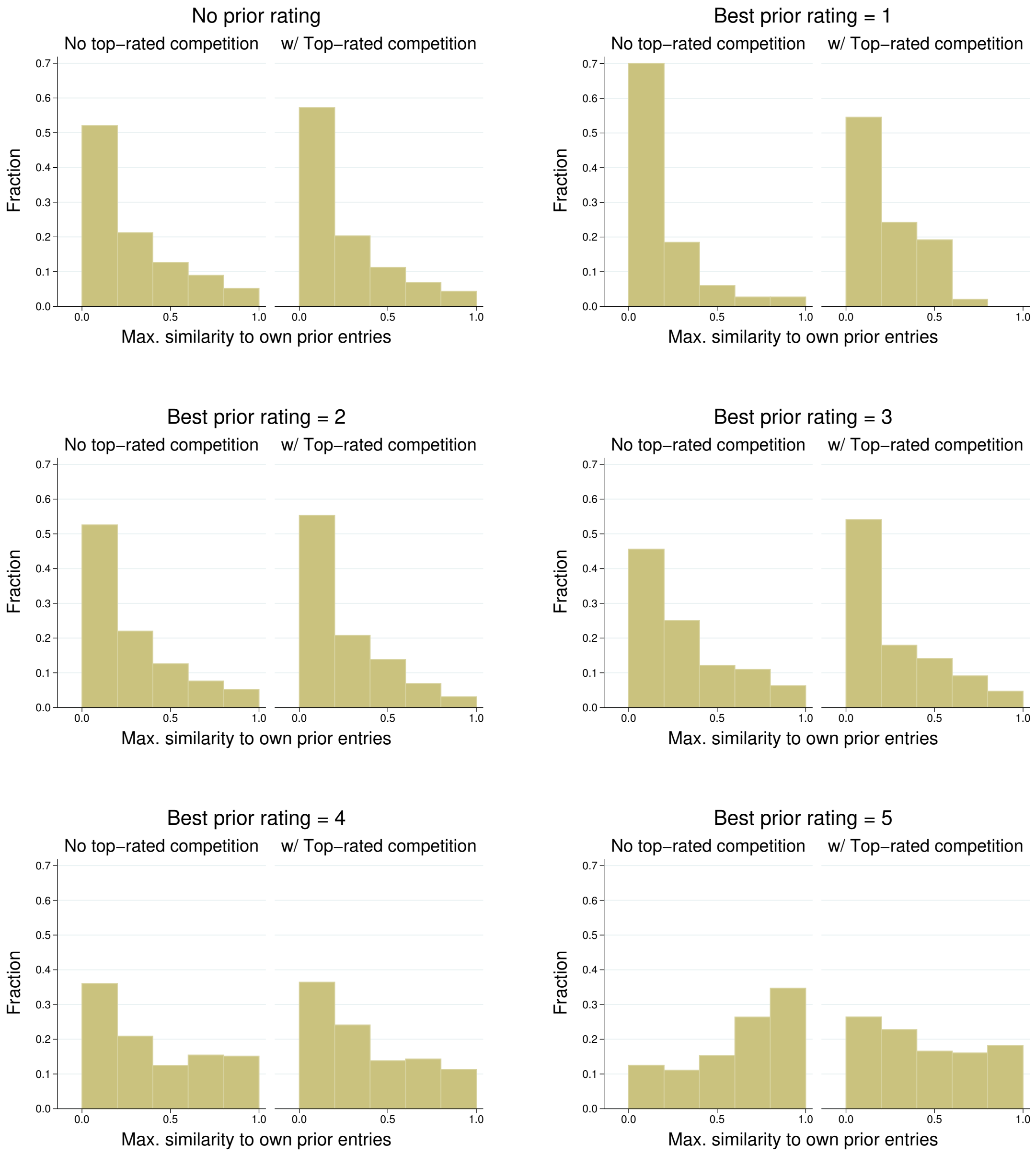
Train, Kenneth E., *Discrete Choice Methods with Simulation*, Cambridge: Cambridge University Press, 2009.

Williams, Heidi, "Innovation Inducement Prizes: Connecting Research to Policy," *Journal of Policy Analysis and Management*, 2012, *31* (3), 752–776.

Wooten, Joel and Karl Ulrich, "The Impact of Visibility in Innovation Tournaments: Evidence from field experiments," 2013. Working Paper.

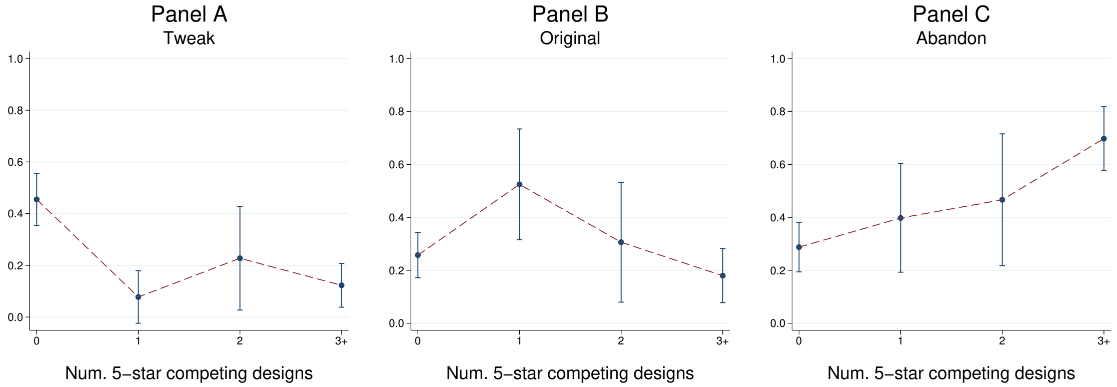
_ **and** _ , "Idea Generation and the Role of Feedback: Evidence from field experiments with innovation tournaments," *Production and Operations Management*, 2017, *26* (1), 80–99.

Figure 1: Distribution of similarity to prior submissions, conditional on ratings and competition



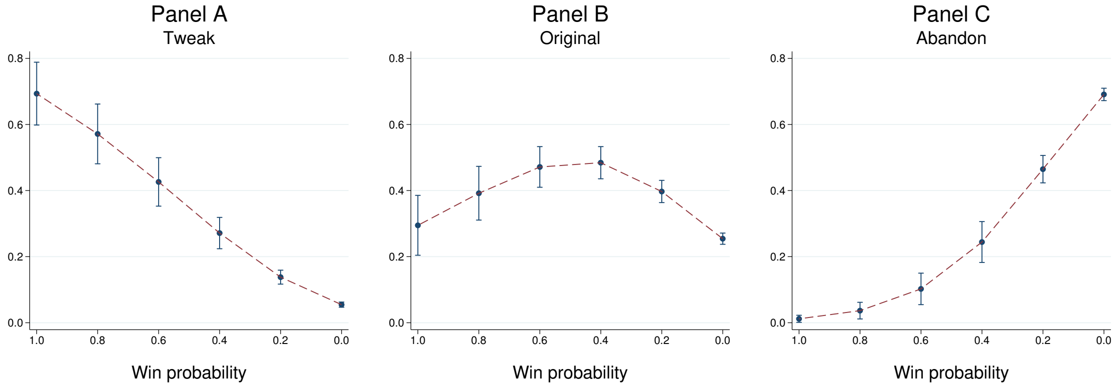
Notes: Figure shows distribution of designs' similarity to prior entries, conditional on the player's highest rating and the presence of top-rated competition. Each observation is a design, and the plotted variable is that design's maximal similarity to any previous submission by the same player in the same contest, taking values in [0,1], where a value of 1 indicates the design is identical to one of the player's earlier submissions.

Figure 2: Probability of tweaks, original designs, and abandonment as a function of 5-star competition



Notes: Figure plots the probability that a player who has at least one 5-star rating in a contest does one of the following on (and after) a given submission: tweaks and then enters more designs (Panel A), experiments and then enters more designs (Panel B), or stops investing in the contest (Panel C). The bars around each point provide the associated 95 percent confidence interval.

Figure 3: Probability of tweaks, original designs, and abandonment as a function of Pr(Win)



Notes: Figure plots the probability that a player does one of the following on (and after) a given submission, as a function of their contemporaneous win probability: tweaks and then enters more designs (Panel A), experiments and then enters more designs (Panel B), or stops investing in the contest (Panel C). These probabilities are estimated as described in the text, and the bars around each point provide the associated 95 percent confidence interval.

Table 1: Characteristics of contests in the sample

Variable	**N**	**Mean**	**SD**	**P25**	**P50**	**P75**
Contest length (days)	122	8.52	3.20	7	7	11
Prize value (US$)	122	247.57	84.92	200	200	225
No. of players	122	33.20	24.46	19	26	39
No. of designs	122	96.38	80.46	52	74	107
5-star designs	122	2.59	4.00	0	1	4
4-star designs	122	12.28	12.13	3	9	18
3-star designs	122	22.16	25.33	6	16	28
2-star designs	122	17.61	25.82	3	10	22
1-star designs	122	12.11	25.24	0	2	11
Unrated designs	122	29.62	31.43	7	19	40
Number rated	122	66.75	71.23	21	50	83
Fraction rated	122	0.64	0.30	0.4	0.7	0.9
Prize committed	122	0.56	0.50	0.0	1.0	1.0
Prize awarded	122	0.85	0.36	1.0	1.0	1.0

Notes: Table reports descriptive statistics for the contests. “Fraction rated” refers to the fraction of designs in each contest that gets rated. “Prize committed” indicates whether the contest prize is committed to be paid (vs. retractable). “Prize awarded” indicates whether the prize was awarded. The fraction of contests awarded awarded subsumes the fraction committed, since committed prizes are always awarded.

Table 2: Distribution of ratings (rated designs only)

	1-star	**2-star**	**3-star**	**4-star**	**5-star**	**Total**
Count	1,478	2,149	2,703	1,498	316	**8,144**
Percent	18.15	26.39	33.19	18.39	3.88	**100**

Notes: Table tabulates rated designs by rating. 69.3 percent of designs in the sample are rated by sponsors on a 1-5 scale. The site provides guidance on the meaning of each rating, which introduces consistency in the interpretation of ratings across contests.

Table 3: Similarity to preceding designs by same player and competitors, and intra-batch similarity

Panel A. Using preferred algorithm: Perceptual Hash

Variable	**N**	**Mean**	**SD**	**P10**	**P50**	**P90**
Max. similarity to any of own preceding designs	5,075	0.32	0.27	0.05	0.22	0.77
Max. similarity to best of own preceding designs	3,871	0.28	0.27	0.03	0.17	0.72
Max. similarity to best of oth. preceding designs	9,709	0.14	0.1	0.04	0.13	0.27
Maximum intra-batch similarity	1,987	0.45	0.32	0.05	0.41	0.91

Panel B. Using alternative algorithm: Difference Hash

Variable	**N**	**Mean**	**SD**	**P10**	**P50**	**P90**
Max. similarity to any of own preceding designs	5,075	0.58	0.28	0.16	0.62	0.94
Max. similarity to best of own preceding designs	3,871	0.52	0.3	0.09	0.54	0.93
Max. similarity to best of oth. preceding designs	9,709	0.33	0.21	0.09	0.29	0.63
Maximum intra-batch similarity	1,987	0.69	0.28	0.23	0.77	0.98

Notes: Table reports summary statistics on designs’ similarity to previously entered designs (both own and competing). Pairwise similarity scores are calculated as described in the text and available for all designs whose digital image could be obtained (96% of entries; refer to the text for an explanation of missing images). The “best” preceding designs are those with the most positive feedback provided prior to the given design. Intra-batch similarity is calculated as the similarity of designs in a given batch to each other, where a design batch is defined to be a set of designs entered by a single player in which each design was entered within 15 minutes of another design in the set. This grouping captures players’ tendency to submit multiple designs at once, which are often similar with minor variations on a theme.

Table 4: Similarity to player's best previously-rated designs

	(1)	(2)	(3)	(4)	(5)
Player's prior best rating==5	0.440***	0.459***	0.260***	0.357***	0.362***
	(0.102)	(0.092)	(0.097)	(0.097)	(0.102)
* 1+ competing 5-stars	-0.197***	-0.245***	-0.158**	-0.206***	-0.208***
	(0.073)	(0.063)	(0.061)	(0.070)	(0.071)
* prize value ($100s)	-0.025	-0.015	0.005	-0.014	-0.018
	(0.025)	(0.023)	(0.032)	(0.031)	(0.033)
Player's prior best rating==4	0.165***	0.160***	0.128***	0.121***	0.116***
	(0.024)	(0.022)	(0.030)	(0.031)	(0.032)
Player's prior best rating==3	0.079***	0.077***	0.068**	0.060**	0.056**
	(0.018)	(0.018)	(0.028)	(0.028)	(0.028)
Player's prior best rating==2	0.044**	0.044**	0.023	0.026	0.024
	(0.021)	(0.022)	(0.029)	(0.030)	(0.030)
One or more competing 5-stars	-0.020	0.009	-0.003	0.004	0.001
	(0.018)	(0.020)	(0.022)	(0.023)	(0.024)
Prize value ($100s)	-0.014*		-0.010		
	(0.007)		(0.010)		
Pct. of contest elapsed	-0.030	-0.060*	-0.010	-0.018	-0.103
	(0.034)	(0.032)	(0.030)	(0.034)	(0.084)
Constant	0.238***	0.207***	0.235***	0.232***	0.303***
	(0.039)	(0.023)	(0.044)	(0.061)	(0.093)
N	3871	3871	3871	3871	3871
R^2	0.07	0.20	0.48	0.53	0.53
Contest FEs	No	Yes	No	Yes	Yes
Player FEs	No	No	Yes	Yes	Yes
Other Controls	No	No	No	No	Yes

Notes: Observations are designs. Dependent variable is a continuous measure of a design's similarity to the highest-rated preceding entry by the same player, taking values in [0,1], where a value of 1 indicates the design is identical to another. The mean value of this variable in the sample is 0.28 (s.d. 0.27). Column (5) controls for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table 5: Change in similarity to player's best previously-rated designs

	(1)	(2)	(3)	(4)	(5)
Δ(Player's best rating==5)	0.861***	0.878***	0.928***	0.914***	0.924***
	(0.162)	(0.170)	(0.203)	(0.205)	(0.205)
* 1+ competing 5-stars	-0.417***	-0.412***	-0.418***	-0.427***	-0.429***
	(0.118)	(0.125)	(0.144)	(0.152)	(0.152)
* prize value ($100s)	-0.092**	-0.094**	-0.115**	-0.107**	-0.110**
	(0.039)	(0.039)	(0.049)	(0.047)	(0.048)
Δ(Player's best rating==4)	0.275***	0.282***	0.267***	0.276***	0.279***
	(0.062)	(0.065)	(0.073)	(0.079)	(0.079)
Δ(Player's best rating==3)	0.143***	0.151***	0.134**	0.137**	0.138**
	(0.055)	(0.058)	(0.065)	(0.069)	(0.069)
Δ(Player's best rating==2)	0.079*	0.082*	0.063	0.059	0.059
	(0.043)	(0.046)	(0.053)	(0.056)	(0.057)
One or more competing 5-stars	-0.003	-0.003	-0.003	0.004	0.003
	(0.007)	(0.015)	(0.014)	(0.025)	(0.026)
Prize value ($100s)	0.003		0.003		
	(0.002)		(0.008)		
Pct. of contest elapsed	0.015	0.009	0.017	0.004	-0.048
	(0.012)	(0.018)	(0.024)	(0.030)	(0.074)
Constant	-0.029***	-0.017*	-0.031	0.063	0.105
	(0.010)	(0.010)	(0.029)	(0.093)	(0.108)
N	2694	2694	2694	2694	2694
R^2	0.03	0.05	0.11	0.14	0.14
Contest FEs	No	Yes	No	Yes	Yes
Player FEs	No	No	Yes	Yes	Yes
Other Controls	No	No	No	No	Yes

Notes: Observations are designs. Dependent variable is a continuous measure of the *change* in designs' similarity to the highest-rated preceding entry by the same player, taking values in [-1,1], where a value of 0 indicates that the player's current design is as similar to her best preceding design as was her previous design, and a value of 1 indicates that the player transitioned fully from innovating to recycling (and a value of -1, the converse). The mean value of this variable in the sample is 0.00 (s.d. 0.23). Column (5) controls for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table 6: Similarity to other designs in the same submission batch

	Unweighted		Weighted	
	(1)	(2)	(3)	(4)
Player's prior best rating==5	0.223	0.238	0.254	0.285
	(0.311)	(0.304)	(0.304)	(0.296)
* 1+ competing 5-stars	-0.308*	-0.305*	-0.303*	-0.295*
	(0.163)	(0.162)	(0.171)	(0.168)
* prize value ($100s)	0.016	0.015	0.010	0.009
	(0.099)	(0.097)	(0.096)	(0.093)
Player's prior best rating==4	0.054*	0.065*	0.064**	0.086**
	(0.032)	(0.037)	(0.032)	(0.038)
Player's prior best rating==3	0.055	0.062*	0.052	0.065*
	(0.035)	(0.037)	(0.035)	(0.037)
Player's prior best rating==2	0.021	0.027	0.007	0.018
	(0.050)	(0.051)	(0.047)	(0.047)
One or more competing 5-stars	0.025	0.027	0.027	0.027
	(0.048)	(0.049)	(0.053)	(0.054)
Pct. of contest elapsed	-0.023	-0.093	-0.010	-0.056
	(0.049)	(0.114)	(0.050)	(0.111)
Constant	0.400***	0.507***	0.391***	0.459***
	(0.066)	(0.148)	(0.060)	(0.146)
N	1987	1987	1987	1987
R^2	0.57	0.57	0.58	0.58
Contest FEs	Yes	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes	Yes
Other Controls	No	Yes	No	Yes

Notes: Observations are design batches, which are defined to be a set of designs by a single player entered into a contest in close proximity (15 minutes). Dependent variable is a continuous measure of intra-batch similarity, taking values in [0,1], where a value of 1 indicates that two designs in the batch are identical. The mean value of this variable in the sample is 0.45 (s.d. 0.32). Columns (3) and (4) weight the regressions by batch size. Columns (2) and (4) control for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table 7: Similarity to player's best not-yet-rated designs (placebo test)

	Similarity to forthcoming			Residual
	(1)	(2)	(3)	(4)
Player's best forthcoming rating==5	0.007	-0.084	-0.105	-0.113
	(0.169)	(0.136)	(0.151)	(0.122)
* 1+ competing 5-stars	-0.094	0.032	0.027	0.035
	(0.099)	(0.056)	(0.066)	(0.062)
* prize value ($100s)	-0.003	0.015	0.021	0.018
	(0.031)	(0.025)	(0.027)	(0.025)
Player's best forthcoming rating==4	0.039	0.051	0.049	0.034
	(0.066)	(0.096)	(0.094)	(0.095)
Player's best forthcoming rating==3	0.080	0.049	0.051	0.036
	(0.052)	(0.088)	(0.088)	(0.088)
Player's best forthcoming rating==2	0.030	-0.010	-0.007	-0.014
	(0.049)	(0.093)	(0.094)	(0.095)
One or more competing 5-stars	-0.080	-0.013	-0.010	-0.013
	(0.097)	(0.110)	(0.117)	(0.119)
Pct. of contest elapsed	0.016	-0.502	-0.466	-0.468
	(0.242)	(0.478)	(0.462)	(0.497)
Constant	0.217	0.556	0.569	0.398
	(0.212)	(0.560)	(0.543)	(0.581)
N	1147	577	577	577
R^2	0.68	0.83	0.83	0.67
Contest FEs	Yes	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes

Notes: Table provides a placebo test of the effects of future feedback on similarity. Observations are designs. Dependent variable in Columns (1) to (3) is a continuous measure of a design's similarity to the best design that the player has previously entered that has yet to *but will eventually be* rated, taking values in [0,1], where a value of 1 indicates that the two designs are identical. The mean value of this variable is 0.26 (s.d. 0.25). Under the identifying assumption that future feedback is unpredictable, current choices should be unrelated to forthcoming ratings. Note that a given design's similarity to an earlier, unrated design can be incidental if they are both tweaks on a rated third design. To account for this possibility, Column (2) controls for the given and unrated designs' similarity to the best previously-rated design. Column (3) allows these controls to vary with the highest rating previously received. Dependent variable in Column (4) is the residual from a regression of the dependent variable in the previous columns on these controls. These residuals will be the subset of a given design's similarity to the unrated design that is not explained by jointly-occurring similarity to a third design. All columns control for days remaining and number of previous designs by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table 8: Similarity to any of player's previous designs: 4-vs-4

	(1)	(2)	(3)
Player's prior best rating==4	0.185***	0.382**	0.564*
	(0.069)	(0.190)	(0.299)
* 1+ competing 4- or 5-stars	-0.121**	-0.323*	-0.315
	(0.051)	(0.181)	(0.248)
* prize value ($100s)	-0.011	-0.006	-0.045
	(0.020)	(0.033)	(0.097)
Player's prior best rating==3	0.006	0.042	-0.029
	(0.023)	(0.041)	(0.088)
Player's prior best rating==2	-0.016	-0.046	0.040
	(0.032)	(0.057)	(0.097)
One or more competing 4- or 5-stars	0.060**	0.076	0.091
	(0.028)	(0.092)	(0.169)
Pct. of contest elapsed	-0.091	-0.638**	0.828
	(0.119)	(0.274)	(0.755)
Constant	0.457**	0.948***	-0.320
	(0.209)	(0.317)	(0.638)
N	2926	1557	879
R^2	0.52	0.60	0.67
Contest FEs	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes
Restriction	All	2nd half	4th qtr

Notes: Table shows the effects of 4-star feedback and competition on similarity when no player has a 5-star rating. Observations are designs. Dependent variable is a continuous measure of a design's maximal similarity to previous entries in the same contest by the same player, taking values in [0,1], where a value of 1 indicates the design is identical to another. All columns include contest and player fixed effects and control for the number of days remaining and number of previous designs entered by the player and her competitors. Columns (2) and (3) restrict the sample to submissions in the second half or fourth quarter of a contest, when the absence of 5-star ratings may be more meaningful and is increasingly likely to be final. Similarity scores in this table are calculated using a perceptual hash algorithm. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table 9: Initial submission similarity to assorted comparison groups

	(1)	(2)	(3)	(4)
Competitors' best rating==5	-0.079**	0.006	-0.003	-0.015
	(0.032)	(0.040)	(0.034)	(0.038)
Competitors' best rating==4	-0.021	0.009	0.005	-0.008
	(0.030)	(0.039)	(0.031)	(0.037)
Competitors' best rating==3	-0.002	-0.008	0.025	0.001
	(0.030)	(0.039)	(0.031)	(0.038)
Competitors' best rating==2	-0.028	-0.002	0.019	0.033
	(0.037)	(0.043)	(0.042)	(0.047)
Pct. of contest elapsed	0.091***	0.122***	0.024	-0.003
	(0.033)	(0.037)	(0.038)	(0.036)
Constant	0.151***	0.185***	0.253***	0.472***
	(0.044)	(0.054)	(0.050)	(0.056)
N	2996	2996	2507	2996
R^2	0.56	0.63	0.74	0.68
Contest FEs	Yes	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes

Notes: Table shows the effects of competition at the time a player makes their first submission in a contest on that submission's similarity to the highest-rated prior design in the contest (Column 1), all prior designs in the contest (Column 2), all other designs by the same player in other contests in the data (Column 3), and all other designs by any player in other contests in the data (Column 4). All columns include contest and player fixed effects and control for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table 10: Latent-utility estimates for each action, as a function of Pr(Win)

	Tweak	**Original**	**Abandon**
Latent utility when	$\boxed{4.065}$	3.210	0.000
Pr(Win) = 100%:	$\boxed{(1.113)}$	(0.775)	n.a.
Latent utility when	$\boxed{2.745}$	2.368	0.000
Pr(Win) = 80%:	$\boxed{(1.011)}$	(0.687)	n.a.
Latent utility when	1.425	$\boxed{1.526}$	0.000
Pr(Win) = 60%:	(0.909)	$\boxed{(0.599)}$	n.a.
Latent utility when	0.105	$\boxed{0.684}$	0.000
Pr(Win) = 40%:	(0.806)	$\boxed{(0.511)}$	n.a.
Latent utility when	-1.215	-0.157	$\boxed{0.000}$
Pr(Win) = 20%:	(0.704)	(0.422)	$\boxed{\text{n.a.}}$
Latent utility when	-2.535	-0.999	$\boxed{0.000}$
Pr(Win) = 0%:	(0.602)	(0.334)	$\boxed{\text{n.a.}}$

Notes: Table shows latent-utility estimates from a choice model relating players' actions to their contemporaneous win probability, evaluated at six values of Pr(Win). These estimates represent the latent utility of each action relative to the outside option of abandonment, which has utility normalized to zero. See text for discussion. Boxes identify action with greatest utility. Robust standard errors in parentheses.

Table 11: Normalized panelist ratings on tweaks vs. original designs

	Outcomes for:		**Diff. in**
Metric	**Tweaks**	**Originals**	**means**
PCA score of	-0.45	0.18	0.64***
panelist ratings	(0.21)	(0.15)	p=0.008
Average rating	-0.45	0.22	0.67***
by panelists	(0.20)	(0.14)	p=0.004
Median rating	-0.46	0.23	0.69***
by panelists	(0.21)	(0.15)	p=0.005
Max rating	1.08	1.99	0.91***
by panelists	(0.22)	(0.17)	p=0.001
Disagreement (s.d.)	1.34	1.59	0.25**
among panelists	(0.10)	(0.07)	p=0.019

Notes: Table compares professional graphic designers' ratings on tweaks and original designs that received a top rating from contest sponsors. Panelists' ratings were demeaned prior to analysis. The PCA score refers to a design's score along the first component from a principal component component analysis of panelists' ratings. The other summary measures are the mean, median, max, and s.d. of panelists' ratings on a given design. A design is classified as a tweak if its maximal similarity to any previous design by that player is greater than 0.7 and as original if it is less than 0.3. Standard errors in parentheses are provided below each mean, and results from a one-sided test of equality of means is provided to the right. *, **, *** indicate significance at the 0.1, 0.05, and 0.01 levels, respectively. Similarity scores calculated using perceptual hash algorithm. Results are robust to both algorithms and alternative cutoffs for originality.

Appendix for Online Publication

A Theoretical Background

Before proving the propositions of the paper, it is useful to establish an identity. Proposition 1 requires that $q \in \left(\frac{1}{1+\alpha}, \frac{1}{2}\right)$. In the text of the paper, it is stated that this condition implies that exploration has higher expected quality than exploitation. To see this, observe that:

$$q > \tfrac{1}{1+\alpha} \implies q > \tfrac{\alpha-1}{\alpha^2-1} \implies q\left(\alpha^2-1\right) - (\alpha-1) > 0 \implies q\left(\alpha - \tfrac{1}{\alpha}\right) - \left(1 - \tfrac{1}{\alpha}\right) > 0$$
$$\implies q\alpha + (1-q)\tfrac{1}{\alpha} > 1 \implies \underbrace{q\alpha\beta_0 + (1-q)\frac{1}{\alpha}\beta_0}_{E[\beta_1|\text{Explore}]} > \underbrace{\beta_0}_{E[\beta_1|\text{Explore}]}$$

Note that here, as in the proofs below, the j subscript is omitted to simplify notation.

Proposition 1: Suppose $q \in \left(\frac{1}{1+\alpha}, \frac{1}{2}\right)$. Then, there exists a μ^* such that for all $\mu_j < \mu^*$,

$$\underbrace{F\left(\beta_{j0}\right)}_{E[Pr(Win)|exploit]} > \underbrace{\left[qF\left(\beta_{j1}^H\right) + (1-q)F\left(\beta_{j1}^L\right)\right]}_{E[Pr(Win)|explore]}$$

and for all $\mu_j > \mu^*$,

$$\underbrace{\left[qF\left(\beta_{j1}^H\right) + (1-q)F\left(\beta_{j1}^L\right)\right]}_{E[Pr(Win)|explore]} > \underbrace{F\left(\beta_{j0}\right)}_{E[Pr(Win)|exploit]}$$

Proof:

To prove this statement, it is sufficient to show that (i) the difference in returns to exploration over exploitation is zero when $\mu = 0$, (ii) the first derivative of this function is negative when $\mu = 0$, and (iii) the function has exactly one positive, real root. These three conditions imply a function that is negative for low μ and positive for high μ. The proof proceeds in sequence.

Proof of (i): When $\mu = 0$,

$$qF\left(\beta_1^H\right) + (1-q)F\left(\beta_1^L\right) - F\left(\beta_0\right)$$
$$= q\left(\frac{(1+\alpha)\beta_0}{(1+\alpha)\beta_0 + 0}\right) + (1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_0}{\left(1+\frac{1}{\alpha}\right)\beta_0 + 0}\right) - \left(\frac{2\beta_0}{2\beta_0 + 0}\right) = q + (1-q) - 1 = 0$$

Proof of (ii): When $\mu = 0$,

$$\begin{aligned}
&\frac{\partial}{\partial \mu}\left[qF\left(\beta_1^H\right) + (1-q)\,F\left(\beta_1^L\right) - F\left(\beta_0\right)\right]_{\mu=0} \\
&\qquad = \frac{\partial}{\partial \mu}\left[q\left(\frac{(1+\alpha)\,\beta_0}{(1+\alpha)\,\beta_0+\mu}\right) + (1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_0}{\left(1+\frac{1}{\alpha}\right)\beta_0+\mu}\right) - \left(\frac{2\beta_0}{2\beta_0+\mu}\right)\right]_{\mu=0} \\
&\qquad = q\left(\frac{-(1+\alpha)\,\beta_0}{\left((1+\alpha)\,\beta_0+0\right)^2}\right) + (1-q)\left(\frac{-\left(1+\frac{1}{\alpha}\right)\beta_0}{\left(\left(1+\frac{1}{\alpha}\right)\beta_0+0\right)^2}\right) + \frac{2\beta_0}{\left(2\beta_0+0\right)^2} \\
&\qquad = q\left(\frac{-1}{(1+\alpha)\,\beta_0}\right) + (1-q)\left(\frac{-1}{\left(1+\frac{1}{\alpha}\right)\beta_0}\right) + \frac{1}{2\beta_0} \\
&\qquad = \frac{1}{2\,(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_0}\cdot\left[q\cdot -2\left(1+\frac{1}{\alpha}\right) + (1-q)\cdot -2\,(1+\alpha) + (1+\alpha)\left(1+\frac{1}{\alpha}\right)\right] \\
&\qquad = \frac{1}{2\alpha\,(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_0}\cdot\left[q\cdot -2\,(1+\alpha) + (1-q)\cdot -2\,(1+\alpha)\,\alpha + (1+\alpha)^2\right] \\
&\qquad = \frac{1}{2\alpha\left(1+\frac{1}{\alpha}\right)\beta_0}\cdot\left[-2q - 2\,(1-q)\,\alpha + (1+\alpha)\right] \\
&\qquad = \frac{1}{2\alpha\left(1+\frac{1}{\alpha}\right)\beta_0}\cdot\left[-2q + 2q\alpha - \alpha + 1\right] \\
&\qquad = \frac{1}{2\alpha\left(1+\frac{1}{\alpha}\right)\beta_0}\cdot\left[2q\,(\alpha-1) - (\alpha-1)\right] \\
&\qquad = \frac{\alpha-1}{2\alpha\left(1+\frac{1}{\alpha}\right)\beta_0}\cdot\left[2q-1\right] < 0 \quad \text{, because } \alpha > 1 \text{ and } q < 1/2
\end{aligned}$$

Proof of (iii): To find the roots, we must solve

$$\begin{aligned}
&qF\left(\beta_1^H\right) + (1-q)\,F\left(\beta_1^L\right) - F\left(\beta_0\right) \\
&\qquad = q\left(\frac{(1+\alpha)\,\beta_0}{(1+\alpha)\,\beta_0+\mu}\right) + (1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_0}{\left(1+\frac{1}{\alpha}\right)\beta_0+\mu}\right) - \left(\frac{2\beta_0}{2\beta_0+\mu}\right) = 0
\end{aligned}$$

To simplify notation, let

$$\begin{aligned}
X &= (1+\alpha)\,\beta_0 \\
Y &= \left(1+\frac{1}{\alpha}\right)\beta_0 \\
Z &= 2\beta_0
\end{aligned}$$

Multiplying out the denominators, we can then rewrite the equation as:

$$qX\,(Y+\mu)\,(Z+\mu) + (1-q)\,Y\,(X+\mu)\,(Z+\mu) - Z\,(X+\mu)\,(Y+\mu) = 0$$

Expanding and combining terms, the equation transforms to the following quadratic:

$$\mu^2\left(qX + (1-q)\,Y - Z\right) + \mu\left(qXZ + XY - qYZ - XZ\right) = 0$$

From this expression, it is immediately apparent that one of the two potential roots is $\mu = 0$ – as found in part (i) of this proof. To find the sign of the nonzero root, we can write:

$$\mu = -\frac{q\left(X - Y\right)Z + X\left(Y - Z\right)}{q\left(X - Y\right) + \left(Y - Z\right)}$$

Reverting back to the original notation,

$$\mu = -\frac{q \cdot \left(\left(1+\alpha\right)\beta_0 - \left(1 + \frac{1}{\alpha}\right)\beta_0\right) \cdot 2\beta_0 + \left(1+\alpha\right)\beta_0 \cdot \left(\left(1 + \frac{1}{\alpha}\right)\beta_0 - 2\beta_0\right)}{q\left(\left(1+\alpha\right)\beta_0 - \left(1 + \frac{1}{\alpha}\right)\beta_0\right) + \left(\left(1 + \frac{1}{\alpha}\right)\beta_0 - 2\beta_0\right)}$$

Which simplifies to:

$$\mu = -\frac{2q\left(\alpha - \frac{1}{\alpha}\right)\beta_0^2 + \left(1+\alpha\right)\left(\frac{1}{\alpha} - 1\right)\beta_0^2}{q\left(\alpha - \frac{1}{\alpha}\right)\beta_0 + \left(\frac{1}{\alpha} - 1\right)\beta_0}$$

Multiplying and dividing by α, we can write:

$$\mu = -\frac{2q\left(\alpha^2 - 1\right)\beta_0^2 - \left(\alpha^2 - 1\right)\beta_0^2}{q\left(\alpha^2 - 1\right)\beta_0 - \left(\alpha - 1\right)\beta_0}$$

Then, canceling out $\left(\alpha - 1\right)\beta_0$, we get:

$$\begin{aligned}\mu &= -\frac{2q\left(1+\alpha\right)\beta_0 - \left(1+\alpha\right)\beta_0}{q\left(1+\alpha\right) - 1} \\ &= \frac{\left(1 - 2q\right)\left(1+\alpha\right)\beta_0}{q\left(1+\alpha\right) - 1} > 0\ ,\end{aligned}$$

because $q < \frac{1}{2}$ (which implies that the numerator is positive) and $q > \frac{1}{1+\alpha}$ (which implies that the denominator is positive). Thus, the remaining root is positive.

Proposition 2: The returns to a player's second design decline to zero as $\mu_j \longrightarrow \infty$.

Proof:

The returns to the second submission can be measured in terms of the increase in win probability that it generates. For any value of β_1 (i.e., for both exploration and exploitation), as $\mu \longrightarrow \infty$:

$$F\left(\beta_1\right) - F\left(\beta_0\right) = \left(\frac{\beta_0 + \beta_1}{\beta_0 + \beta_1 + \mu}\right) - \left(\frac{\beta_0}{\beta_0 + \mu}\right) \longrightarrow \left(\frac{\beta_0 + \beta_1}{\mu}\right) - \left(\frac{\beta_0}{\mu}\right) \longrightarrow \left(\frac{\beta_1}{\mu}\right) \longrightarrow 0$$

Extension of results to cases where $d > c$

Although the paper assumes $d = c$, it also notes that the theoretical result that exploration is incentivized at intermediate values of μ is general to cases where $d > c$, as long as exploration is not *prohibitively* costly (i.e., as long as there exists any $\mu \geq 0$ at which exploration is preferred to the alternatives). Provided this condition is met, exploration will be the most profitable choice in some *intermediate interval* $[\mu_1, \mu_2]$, with exploitation preferred for most $\mu < \mu_1$ (except for μ near zero, where abandonment is preferred, as the player is so far ahead that she is all but guaranteed to win), and abandonment preferred for all $\mu > \mu_2$ (as the player is so far behind that the second submission isn't worth her effort).

The result can be obtained through a sequence of four propositions:

Proposition A.1. *Exploration vs. exploitation*
When $q \in \left(\frac{1}{1+\alpha}, \frac{1}{2}\right)$, there exists a unique level of competition μ_2^* at which the payoffs to exploration, relative to exploitation, are maximized.

Proposition A.2. *Exploration vs. abandonment*
For all values of q, there exists a unique level of competition $\mu_1^* < \mu_2^*$ at which the payoffs to exploration, relative to abandonment, are maximized.

Proposition A.3. *Binding constraints*
At very low and very high μ, the next-best alternative to exploration is abandonment. At intermediate μ, the next-best option is exploitation.

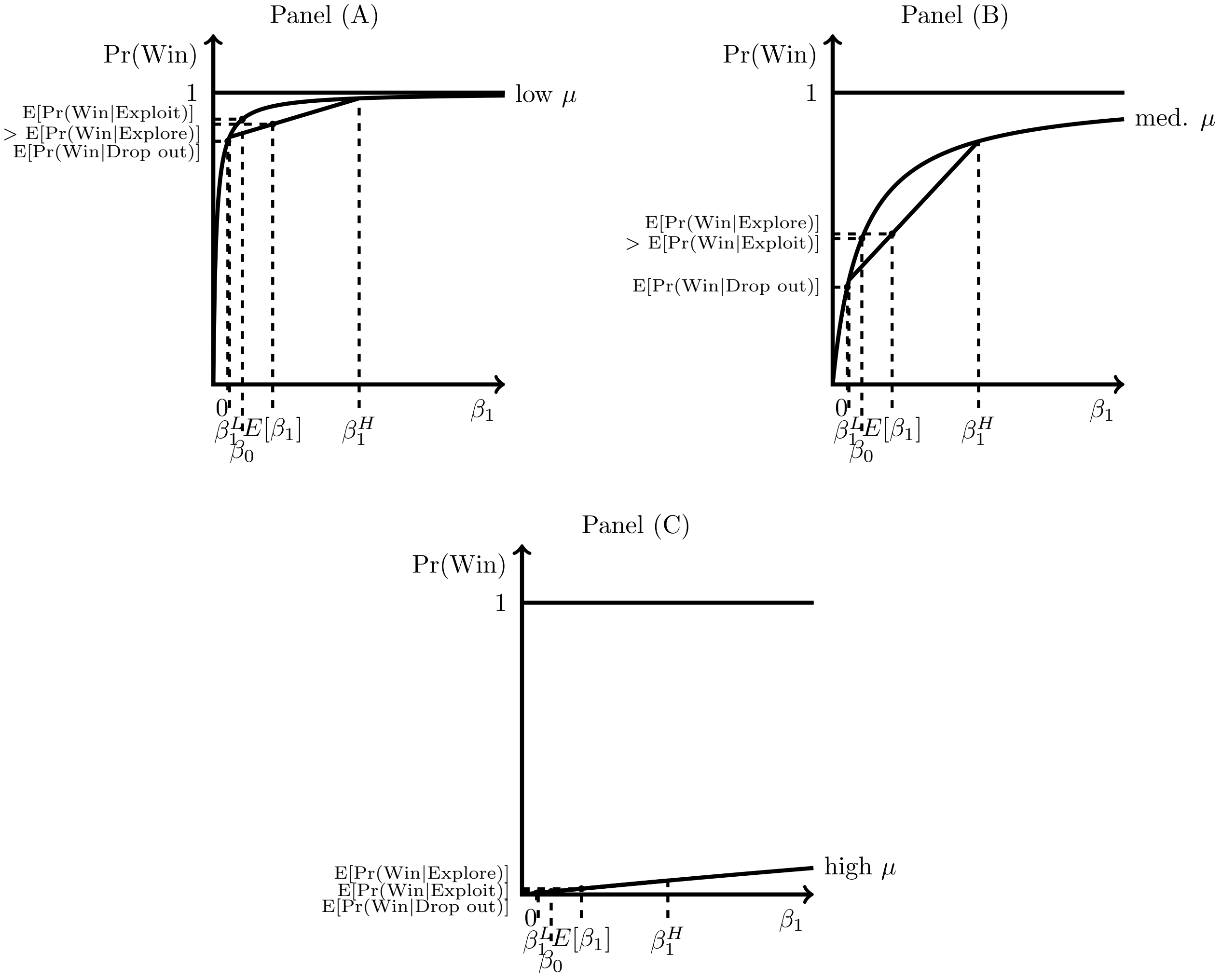
Proposition A.4. *Tying it all together*
When $q \in \left(\frac{1}{1+\alpha}, \frac{1}{2}\right)$, there exists a unique level of competition $\mu^* \in [\mu_1^*, \mu_2^*]$ at which the payoffs to exploration are maximized relative to the player's *next-best* alternative.

Much like the propositions in the body of the paper, these propositions all derive from the shape of the difference in the returns to exploration vs. alternative actions, and they can be shown in a similar way to the proofs above.[1] At a more primitive level, as the paper notes (Footnore 4), the theoretical results are driven by the curvature of the success function, which rises and then flattens with competition – and only at intermediate levels of competition does the function have the curvature to make the returns to exploration both larger than those to exploration and large enough to exceed the cost, as in Figure A.1 below.

Proposition A.4 tells us that the difference in the returns to exploration over its alternatives are greatest at some positive, finite, intermediate value of μ, but it does not guarantee that the difference is in fact greater than zero. This is where the added condition (that there exists some $\mu > 0$ at which exploration *is* a preferred action) comes in, ensuring that this will be the case. Proposition A.4 is illustrated for an example parametrization in Figure A.2 below, which plots the difference in returns to exploration over exploitation, abandonment, and the greater of the two.

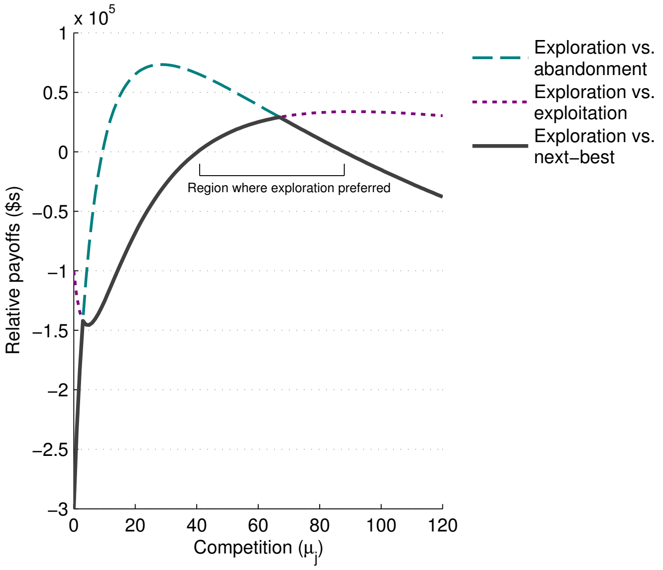
[1] These propositions and proofs were included in an earlier version of this paper and are available from the author.

Figure A.1: Illustration of success function as μ increases



Notes: Figure illustrates expected benefits to exploration, exploitation, and abandonment under low competition (panel A), moderate competition (panel B), and severe competition (panel C). Each subfigure plots a player's probability of winning conditional on the level of competition (μ), the quality of her first design (β_0), and the action taken. The horizontal axis measures the quality of the player's second design. In Panel A, exploitation is preferred to exploration; in panel B, exploration is preferred to exploitation; and in panel C, neither has any significant benefit over abandonment. The gains to exploration are determined by the concavity of the success function in the vicinity of β_0.

Figure A.2: Difference in returns to exploration over alternatives (example)



Notes: Figure plots the incremental payoff to exploration over the next-best alternative as a function of μ_j, for the following parametrization:

$\beta_{j1} = 10$, $q = 0.33$, $\alpha = 9$, $c = 200\text{k}$, $d = 300\text{k}$, $P = 2\text{m}$.

B Dataset Construction

Data were collected on all logo design contests with open (i.e., public) bidding that launched the week of September 3 to 9, 2013, and every three weeks thereafter through the week of November 5 to 11, 2013. Conditional on open bidding, this sample is effectively randomly drawn. The sample used in the paper is further restricted to contests with a single, winner-take-all prize and with no mid-contest rule changes such as prize increases, deadline extensions, and early endings. The sample also excludes one contest that went dormant and resumed after several weeks, as well as a handful of contests whose sponsors simply stopped participating and were never heard from again. These restrictions cause 146 contests to be dropped from the sample. The final dataset includes 122 contests, 4,050 contest-players, and 11,758 designs.[2]

To collect the data, I developed an automated script to scan these contests once an hour for new submissions, save a copy of each design for analysis, and record their owners' identity and performance history from a player profile. I successfully obtained the image files for 96 percent of designs in the final sample. The remaining designs were entered and withdrawn before they could be observed (recall that players can withdraw designs they have entered into a contest, though this option is rarely exercised and can be reversed at the request of a sponsor). All other data were automatically acquired at the conclusion of each contest, once the prize was awarded or the sponsor exercised its outside option of a refund.

B.1 Variables

The dataset includes information on the characteristics of contests, contest-players, and designs:

- Contest-level variables include: the contest sponsor, features of the project brief (title, description, sponsor industry, materials to be included in logo), start and end dates, the prize amount (and whether committed), and the number of players and designs of each rating.
- Contest-player-level variables include: the player's self-reported country, his/her experience in previous contests on the platform (number of contests and designs entered, contests won), and that player's participation and performance in the given contest.
- Design-level variables include: the design's owner, its submission time and order of entry, the feedback it received, the time at which this feedback was given, and whether it was eventually withdrawn. For designs with images acquired, I calculate similarity using the procedures described in the next section. The majority of the analysis occurs at the design level.

Note that designs are occasionally re-rated: five percent of all rated designs are re-rated an average of 1.2 times each. Of these, 14 percent are given their original rating, and 83 percent are re-rated within 1 star of the original rating. I treat the first rating on each design to be the most informative, objective measure of quality, since research suggests first instincts tend to be most reliable and ratings revisions are likely made relative to other designs in the contest rather than an objective benchmark.

[2] While 268 contests were originally sampled, only 122 of these survived the filters described above. The number of contests sampled was constrained by high costs of data collection, which was performed in real-time by making round-the-clock hourly scans for new activity without putting a heavy strain on the platform's servers.

B.2 Image Comparison Algorithms

This paper uses two distinct algorithms to calculate pairwise similarity scores. One is a perceptual hash algorithm, which creates a digital signature (hash) for each image from its lowest frequency content. As the name implies, a perceptual hash is designed to imitate human perception. The second algorithm is a difference hash, which creates the hash from pixel intensity gradients.

I implement the perceptual hash algorithm and calculate pairwise similarity scores using a variant of the procedure described by the Hacker Factor blog.[3] This requires six steps:

1. Resize each image to 32x32 pixels and convert to grayscale.
2. Compute the discrete cosine transform (DCT) of each image. The DCT is a widely-used transform in signal processing that expresses a finite sequence of data points as a linear combination of cosine functions oscillating at different frequencies. By isolating low frequency content, the DCT reduces a signal (in this case, an image) to its underlying structure. The DCT is broadly used in digital media compression, including MP3 and JPEG formats.
3. Retain the upper-left 16x16 DCT coefficients and calculate the average value, excluding first term.
4. Assign 1s to grid cells with above-average DCT coefficients, and 0s elsewhere.
5. Reshape to 256 bit string; this is the image's digital signature (hash).
6. Compute the Hamming distance between the two hashes and divide by 256.

The similarity score is obtained by subtracting this fraction from one. In a series of sensitivity tests, the perceptual hash algorithm was found to be strongly invariant to transformations in scale, aspect ratio, brightness, and contrast, albeit not rotation. As described, the algorithm will perceive two images that have inverted colors but are otherwise identical to be perfectly dissimilar. I make the algorithm robust to color inversion by comparing each image against the regular and inverted hash of its counterpart in the pair, taking the maximum similarity score, and rescaling so that the scores remain in [0,1]. The resulting score is approximately the absolute value correlation of two images' content.

I follow a similar procedure outlined by the same blog[4] to implement the difference hash algorithm and calculate an alternative set of similarity scores for robustness checks:

1. Resize each image to 17x16 pixels and convert to grayscale.
2. Calculate horizontal gradient as the change in pixel intensity from left to right, returning a 16x16 grid (note: top to bottom is an equally valid alternative)
3. Assign 1s to grid cells with positive gradient, 0s to cells with negative gradient.
4. Reshape to 256 bit string; this is the image's digital signature (hash).
5. Compute the Hamming distance between the two hashes and divide by 256.

[3] See http://www.hackerfactor.com/blog/archives/432-Looks-Like-It.html.

[4] See http://www.hackerfactor.com/blog/archives/529-Kind-of-Like-That.html.

The similarity score is obtained by subtracting this fraction from one. In sensitivity tests, the difference hash algorithm was found to be highly invariant to transformations in scale and aspect ratio, potentially sensitive to changes in brightness and contrast, and very sensitive to rotation. I make the algorithm robust to color inversion using a procedure identical to that described for the perceptual hash.

Though the perceptual and difference hash algorithms are both conceptually and mathematically distinct, and the resulting similarity scores are only modestly correlated ($\rho = 0.38$), the empirical results of Section 3 are qualitatively and quantitatively similar under either algorithm. This consistency is reassurance that the patterns found are not simply an artifact of an arcane image processing algorithm; rather, they appear to be generated by the visual content of the images themselves.

B.3 Why use algorithms?

There are two advantages to using algorithms over human judges. The first is that the algorithms can be directed to evaluate specific features of an image and thus provide a consistent, objective measure of similarity, whereas individuals may be attuned to different features and can have different perceptions of similarity in practice (Tirilly et al. 2012). This argument is supported by a pilot study I attempted using Amazon Mechanical Turk, in which I asked participants to rate the similarity of pairs of images they were shown; the results (not provided here) were generally very noisy, except in cases of nearly identical images, in which case the respondents tended to agree that they were similar. The second advantage of algorithms is more obvious: they are cheap, taking only seconds to execute a comparison.

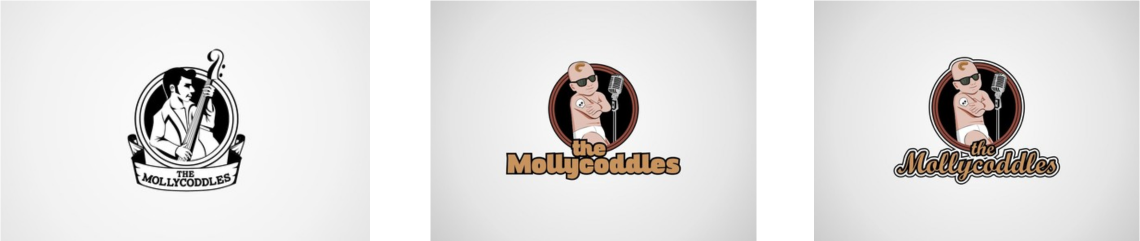
The obvious concern is that computer vision may not be comparable to human perception, in which case the measures are not relevant to human behavior. This concern is mitigated by the fact that they are used to approximate human perception in commercial software, as well as two pieces of evidence from the above studies: (i) when two images are similar, humans and algorithms tend to agree, and (ii) when two images are dissimilar, neither humans nor algorithms find them similar, but they may also disagree on the degree of dissimilarity. Thus, human and computer judgment tend to align within coarse categories, especially at extremes – which is the margin of variation that matters most for this paper.

The evidence of disagreement in subjects' assessments of similarity nevertheless raises a deeper question: is it sensible to apply a uniform similarity measure at all? Squire and Pun (1997) find that *expert* subjects' assessments of similarity tend to agree at all levels, and the designers in this paper could reasonably be classified as visual experts. Even so, it is reassuring that the results throughout this paper are similar in sign, significance, and magnitude under two fundamentally different algorithms, and thus emerge no matter which features we choose to focus on when measuring similarity.

B.4 How do the algorithms perform?

In my own experience browsing the designs in the dataset, images that look similar to the naked eye tend to have a high similarity score, particularly under the perceptual hash algorithm. But as Tirilly et al. (2012) show, similarity is in the eye of the beholder – particularly at intermediate levels and when it is being assessed by laypersons. Figure B.1 illustrates the performance of the algorithms for three logos entered in the order shown by the same player in one contest (not necessarily from the sampled platform):

Figure B.1: Performance of image comparison algorithms



(1) (2) (3)

Notes: Figure shows three logos entered in order by a single player in a single contest. The perceptual hash algorithm calculates a similarity score of 0.313 for logos (1) and (2) and a score of 0.711 for (2) and (3). The difference hash algorithm calculates similarity scores of 0.508 for (1) and (2) and 0.891 for (2) and (3).

The first two images have several features in common but also have some notable differences. Each is centered, defined by a circular frame with text underneath, and presented against a similar backdrop. However the content of the circular frame and the font of the text below are considerably different, and the first logo is in black and white while the second one is in color. The perceptual hash algorithm assigns these two logos a similarity score of 31 percent, while the difference hash gives them 51 percent.

In contrast, the second two images appear much more similar. They again have similar layouts, but now they share the same color assignments and the same content in the frame. Lesser differences remain, primarily with respect to the font style, but the logos appear broadly similar. The perceptual hash algorithm assigns these two logos a similarity score of 71 percent; the difference hash, 89 percent.

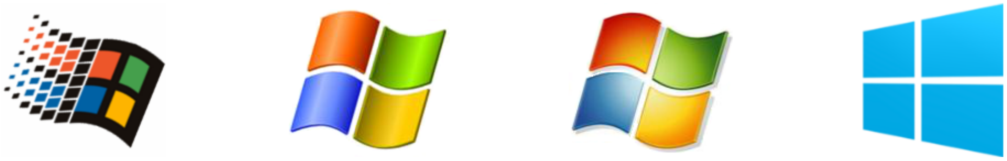
The algorithms thus pass the gut check in this example, which is not particularly unique: further examples using better-known brands are provided below. In light of this evidence, and the consistency of the paper's results, I believe that these algorithms provide empirically valid measures of similarity.

Figure B.2: Volkswagen logo in 1937, 1967, 1995, 1999

(1937) (1967) (1995) (1999)

Notes: Figure shows the evolution of Volkswagen logos since 1937. The perceptual hash algorithm calculates similarity scores of 0.055 for the 1937 and 1967 logos, 0.430 for the 1967 and 1995 logos, and 0.844 for the 1995 and 1999 logos. The difference hash algorithm calculates similarity scores of 0.195, 0.539, and 0.953, respectively.

Figure B.3: Microsoft Windows 95, XP, 7, and 8 logos



(Windows 95) (Windows XP) (Windows 7) (Windows 8)

Notes: Figure shows a sequence of Windows logos. The perceptual hash algorithm calculates similarity scores of 0.195 for the Windows 95 and XP logos, 0.531 for the Windows XP and 7 logos, and 0.148 for the Windows 7 and 8 logos. The difference hash algorithm calculates similarity scores of 0.055, 0.563, and 0.117, respectively. The reason why the similarity of the Windows XP and 7 logos is not evaluated to be even higher is because the contrast generated by the latter's spotlight and shadow changes the structure of the image (for example, it changes the intensity gradient calculated by the difference hash algorithm).

Appendix References:

[1] Tirilly, Pierre, Chunsheng Huang, Wooseob Jeong, Xiangming Mu, Iris Xie, and Jin Zhang. 2012. "Image Similarity as Assessed by Users: A Quantitative Study." *Proceedings of the American Society for Information Science and Technology*, 49(1), pp. 1-10.

[2] Squire, David and Thierry Pun. 1997. "A Comparison of Human and Machine Assessments of Image Similarity for the Organization of Image Databases." *Proceedings of the Scandinavian Conference on Image Analysis*, Lappeenranta, Finland.

C Additional Contest Characteristics

To highlight some of the basic relationships in the contests on this platform, I reproduce a subset of results from Gross (2017), which studies a larger sample from the same setting. Table C.1 estimates the relationship between contest characteristics such as the prize or frequency of feedback and key outcomes, and Table C.2 estimates the relationship between a design's rating and its probability of being selected.

Correlation of contest characteristics and outcomes

The estimates in Table C.1 suggest that an extra $100 in prize value on average attracts an additional 13.3 players, 47.7 designs, and 0.1 designs per player and increases the odds that a retractable prize will be awarded by 1.6 percent at the mean of all covariates. There is only a modest incremental effect of committed prize dollars, likely because the vast majority of uncommitted prizes are awarded anyway. The effects of feedback are also powerful: a sponsor who rates a high fraction of the designs in the contest will typically see fewer players enter but receive more designs from the participating players and have a much higher probability of finding a design it likes enough to award the prize. The effect of full feedback (relative to no feedback) on the probability the prize is awarded is greater than that of a $1000 increase in the prize – a more than quadrupling of the average and median prize in the sample.

Table C.1: Correlations of contest outcomes with their characteristics

	(1)	(2)	(3)	(4)
	Players	Designs	Designs/Player	Awarded
Total Prize Value ($100s)	13.314***	47.695***	0.050***	0.101***
	(0.713)	(2.930)	(0.016)	(0.027)
Committed Value ($100s)	2.469**	8.674*	0.038	
	(1.205)	(4.932)	(0.025)	
Fraction Rated	-7.321***	15.195***	1.026***	1.121***
	(0.813)	(3.098)	(0.041)	(0.102)
Contest Length	0.537***	2.109***	0.013***	0.021**
	(0.073)	(0.283)	(0.004)	(0.010)
Words in Desc. (100s)	0.130	3.228***	0.063***	-0.143***
	(0.092)	(0.449)	(0.006)	(0.013)
Attached Materials	-0.943***	-1.884***	0.048***	-0.021
	(0.173)	(0.692)	(0.013)	(0.015)
Prize Committed	1.398	4.539	-0.007	
	(3.559)	(14.552)	(0.087)	
Constant	-2.445	-63.730***	1.916***	1.085***
	(2.309)	(9.045)	(0.072)	(0.155)
N	4294	4294	4294	3298
R^2	0.57	0.54	0.22	

Notes: Table shows the estimated effect of contest attributes on overall participation and the probability that the prize is awarded, using the larger sample of Gross (2017). The final specification is estimated as a probit on contests without a committed prize. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Monthly fixed effects included but not shown. Robust SEs in parentheses.

Estimationg the success function: details

Recall the following model from the text: let R_{ijk} denote the rating on design i by player j in contest k, and (in a slight abuse of notation) let $R_{ijk} = \emptyset$ when design ijk is unrated. The value of each design, ν_{ijk}, can then be written as follows:

$$\nu_{ijk} = \gamma_{\emptyset}\mathbb{1}(R_{ijk} = \emptyset) + \gamma_1\mathbb{1}(R_{ijk} = 1) + \ldots + \gamma_5\mathbb{1}(R_{ijk} = 5) + \varepsilon_{ijk} \equiv \psi_{ijk} + \varepsilon_{ijk} \tag{3}$$

As in the theoretical model, the sponsor is assumed to select as winner the design with the highest value. In estimating the γ parameters, each sponsor's choice set of designs is assumed to satisfy I.I.A.; in principle, the submission of a design of any rating in a given contest will reduce competing designs' chances of winning proportionally. For contests with an uncommitted prize, the choice set also includes an outside option of not awarding the prize, with value normalized to zero. Letting I_{jk} be the set of designs by player j in contest k, and I_k be the set of all designs in contest k, player jk's probability of winning is:

$$Pr(j \text{ wins } k) = \frac{\sum_{i \in I_{jk}} e^{\psi_{ijk}}}{\sum_{i \in I_k} e^{\psi_{ik}} + \mathbb{1}(\text{Uncommitted prize})}$$

I use the sample of 496,401 designs in 4,294 contests from Gross (2017) to estimate this model by maximum likelihood. The results are reproduced in Table C.2.

Table C.2: Conditional logit of win-lose outcomes on ratings

Model: Latent design value $\nu_{ijk} = \gamma_5 + \gamma_4 + \gamma_3 + \gamma_2 + \gamma_1 + \gamma_{\emptyset} + \varepsilon_{ijk}$

Fixed effect	**Est.**	**S.E.**	**t-stat**	**Implied β (Appendix A)**
Rating==5	1.53	0.07	22.17	4.618
Rating==4	-0.96	0.06	-15.35	0.383
Rating==3	-3.39	0.08	-40.01	0.034
Rating==2	-5.20	0.17	-30.16	0.006
Rating==1	-6.02	0.28	-21.82	0.002
No rating	-3.43	0.06	-55.35	0.032

Notes: Table provides results from a conditional logit estimation of the win-lose outcome of each design as a function of its rating, using the larger sample of Gross (2017). Outside option is not awarding the prize, with utility normalized to zero. The results can be used to approximate a player's probability of winning a contest as a function of her ratings. As a measure of fit, the design predicted by this model as the odds-on favorite wins roughly 50 percent of contests. See Gross (2017) for further discussion.

Table C.3 below sheds more light on the source of the conditional logit estimates, which are difficult to interpret directly. The table shows a cross-tabulation of contests, by the highest rating granted (columns) and the rating of the winning design (rows). The table shows that sponsors typically select the highest-rated design as winner, especially when the highest rating is 4- or 5-stars, but sponsors also often select unrated designs or the outside option. Rarely are 1- or 2-star designs ever awarded.

Table C.3: Frequency of contests, by highest rating and winning rating

Rating of winner	*Highest rating in contest* **Unrated**	**1-star**	**2-star**	**3-star**	**4-star**	**5-star**	**Total**
Not awarded	66	4	12	92	202	85	461
Unrated	142	5	10	59	347	276	839
1-star	.	.	.	3	6	5	14
2-star	.	.	3	11	16	8	38
3-star	.	.	.	43	146	53	242
4-star	.	.	.	.	836	379	1,215
5-star	.	.	.	.	.	1,485	1,485
Total	208	9	25	208	1,553	2,291	4,294

Notes: Table shows the frequency of contests in the Gross (2017) sample by the highest rating granted and the rating of the winning design.

Evidence that the samples are comparable

The dataset in Gross (2017) consists of nearly all logo design contests with open bidding completed on the platform between July 1, 2010 and June 30, 2012, excluding those with zero prizes, multiple prizes, mid-contest rule changes, or otherwise unusual behavior, and it includes nearly all of the same information as the sample in this paper – except for the designs themselves. Although this sample comes from a slightly earlier time period than the one in the present paper (which was collected in the fall of 2013), both cover periods well after the platform was created and its growth had begun to stabilize.

Table C.4 compares characteristics of contests in the two samples. The contests in the Gross (2017) sample period are on average slightly longer, offer larger prizes, and attract a bit more participation relative to the sample of the present paper, but otherwise, the two samples are similar on observables. These differences are mostly due to the presence of a handful of outlying large contests in the Gross (2017) data. Interestingly, although the total number of designs is on average higher in the Gross (2017) sample, the number of designs of each rating is on average the same; the difference in total designs is fully accounted for by an increase in unrated entries. The most notable difference between the two samples is in the fraction of contests with a committed prize (23 percent vs. 56 percent). This discrepancy is explained by the fact that prize commitment only became an option on the platform halfway through the Gross (2017) sample period. Interestingly, the fraction of contests awarded is nevertheless nearly the same in these two samples.

Tables C.5 and C.6 compare the distribution of ratings and batches in the two samples. The tables demonstrate that individual behavior is consistent across samples: sponsors assign each rating, and players enter designs, at roughly the same frequency. The main differences between the two samples are thus isolated to a handful of the overall contest characteristics highlighted in Table C.4.

Table C.4: Comparing Samples: Contest characteristics

	Gross (2016)	**This paper**
Sample size	*4,294*	*122*
Contest length (days)	9.15	8.52
Prize value (US$)	295.22	247.57
No. of players	37.28	33.20
No. of designs	115.52	96.38
5-star designs	3.41	2.59
4-star designs	13.84	12.28
3-star designs	22.16	22.16
2-star designs	16.04	17.61
1-star designs	10.94	12.11
Unrated designs	49.14	29.62
Number rated	66.38	66.75
Fraction rated	0.56	0.64
Prize committed	0.23	0.56
Prize awarded	0.89	0.85

Table C.5: Comparing Samples: Distribution of ratings

	Gross (2016)	**This paper**
Sample size	*285,052*	*8,144*
1 star (in percent)	16.48	18.15
2 stars	24.16	26.39
3 stars	33.38	33.19
4 stars	20.84	18.39
5 stars	5.13	3.88
	100.00	100.00

Table C.6: Comparing Samples: Design batches by size of batch

	Gross (2016)	**This paper**
Sample size	*335,016*	*8,072*
1 design (in percent)	72.46	71.84
2 designs	17.04	18.62
3 designs	5.75	5.57
4 designs	2.50	2.19
5+ designs	2.25	1.77
	100.00	100.00

D Additional Support for Identification

D.1 Timing of feedback unrelated to rating granted

This subsection provides further support for the empirical strategy of estimating the effects of information about relative performance on individual choices. The paper shows that players do not behave in any ways consistent with their being able to forecast forthcoming feedback: behavior is uncorrelated with forthcoming (future) ratings on their previous entries. But we can also show that the timing of feedback is itself difficult to predict, and that it offers no information on performance.

To do so, in Table D.1 I take the subsample of all rated designs and regress the lag between the time of submission and time of feedback on indicators for the rating granted. The dependent variable in Column (1) is this lag, in hours; Column (2), the lag as a percent of contest duration; and Column (3), an indicator for whether the design was rated before the contest ended. All specifications account for contest and player fixed effects, include the standard controls used in the paper, and cluster standard errors by contest. Across all specifications, I find no relationship between feedback lags and ratings.

Table D.1: Correlation of feedback lags with rating granted

	(1)	(2)	(3)
	Lag (hours)	Lag (pct. of contest)	Rated before end?
Rating==5	1.473	0.007	-0.022
	(3.461)	(0.016)	(0.031)
Rating==4	-2.453	-0.013*	0.010
	(1.792)	(0.008)	(0.020)
Rating==3	-0.759	-0.004	0.007
	(2.131)	(0.009)	(0.015)
Rating==2	1.130	0.005	0.006
	(1.922)	(0.008)	(0.011)
Constant	22.779**	0.242***	1.353***
	(11.009)	(0.066)	(0.122)
N	7388	7388	8144
R^2	0.45	0.48	0.63
Contest FEs	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes

Notes: Table illustrates tendency for designs of different ratings to be rated more or less quickly. The results suggest that sponsors are not quicker to rate their favorite designs. Dependent variable in Column (1) is the lag between submission and feedback, in hours; Column (2), this lag as a fraction of contest length; and Column (3), an indicator for whether a design receives feedback before the contest ends. All columns control for the time of entry, the number of previous designs entered by the given player and competitors, and contest and player fixed effects. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by contest in parentheses.

D.2 Written comments are rare

A distinct threat to identification arises if ratings are accompanied by unobserved, written feedback, and these comments provide explicit instruction that generates the patterns found in this paper.

To evaluate this possibility, I draw a new sample of contests from this platform in which written feedback was visible to the public, seemingly by glitch/error. This sample consists of contests from early in the platform's history (significantly preceding the sample period for this paper), but it can nevertheless shed light on the frequency and potential confounding effects of written comments.

Within this sample, sponsors provided written comments to fewer than 8 percent of submissions, though as Table D.2 shows, the frequency is substantially higher for designs rated 4- or 5-stars (20 percent) than for those with poor ratings. Comments take a range of flavors, with many echoing the rating given (e.g., "This is on the right track" or "Not what I'm looking for"), but some make a more explicit request or suggestion of content changes. Because the latter present the risk of a confound, I had individuals read every comment in this sample and determine whether the sponsor suggested specific changes.

Table D.2: Typical distribution of ratings on designs receiving comments

	All designs	**w/ Comments**		**w/ Instructive comments**		
	in sample	**Number**	**% of all**	**Number**	**% of all**	**% of comm.**
Rating	**(1)**	**(2)**	**(2)/(1)**	**(3)**	**(3)/(1)**	**(3)/(2)**
5 stars	719	141	19.61	90	12.52	63.83
4 stars	3,045	551	18.10	401	13.17	72.78
3 stars	5,334	762	14.29	553	10.37	72.57
2 stars	5,205	519	9.97	369	7.09	71.10
1 star	5,114	498	9.74	262	5.12	52.61
Unrated	25,753	1,066	4.14	665	2.58	62.38
Total	**45,170**	**3,537**	**7.83**	**2,340**	**5.18**	**66.16**

Notes: Table tabulates rated designs by rating, within the sample of designs for which written feedback was accessible. The sample was compiled under a separate data collection effort and comprises contests from early in the platform's history (significantly preceding the sample period for this paper), but it is nonetheless informative. The table provides the frequency of each rating within this sample as a whole, then among: (i) designs receiving written comments, and (ii) those receiving written comments deemed instructive. Though higher-rated designs are more likely to receive written comments, the risk of a confounding effect arises only if these comments instruct players to make specific changes. The table shows that instructive comments are disproportionately given to designs with middling ratings – where this commentary may prove most constructive. The net effect is that 5-star designs receive instructive comments at roughly the same rate as 4- and 3-star designs.

On average, roughly two-thirds of comments are instructive in this way. Among designs receiving any comment, the ones most likely to receive an instructive comment are those with middling ratings: ideas that are incomplete but can be improved. As a result, on net we find that 5-, 4-, and 3-star designs all receive instructive comments at comparable rates (10-13 percent, see Table D.2).

Table D.3 formalizes this result with a regression, which provides standard errors. Column (1) regresses an indicator for whether a design in this sample received written feedback on indicators for its rating. Column (2) replaces the dependent variable with an indicator for a design receiving an instructive comment; Column (3) repeats this model for the subsample of designs that received any comment. In all cases, standard errors are clustered by contest. With standard errors in hand, it can be seen that 5-, 4-, and 3-star designs receive instructive comments at statistically similar rates (Table D.3, Column 2).

Table D.3: Raw frequencies of written feedback, by rating

	Commented on (1)	Instructive (2)	Instructive (3)
Rated 5-stars	0.196***	0.125***	0.638***
	(0.026)	(0.018)	(0.049)
Rated 4-stars	0.181***	0.132***	0.728***
	(0.014)	(0.011)	(0.025)
Rated 3-stars	0.143***	0.104***	0.726***
	(0.013)	(0.010)	(0.023)
Rated 2-stars	0.100***	0.071***	0.711***
	(0.013)	(0.010)	(0.024)
Rated 1-star	0.097***	0.051***	0.526***
	(0.019)	(0.010)	(0.029)
Unrated	0.041***	0.026***	0.624***
	(0.005)	(0.003)	(0.040)
N	45170	45170	3537
R^2	0.11	0.08	0.67
Sample	All designs	All designs	Commented

Notes: Table provides results from regressing an indicator for whether a design received any comment (Column 1) or received a constructive comment (Columns 2 and 3) on indicators for its rating. Column (3) restricts the sample to only designs that received any comment. The table reproduces the tabulations in Table D.2, but provides standard errors. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by contest in parentheses.

Given that only around 10 percent of 5-star designs receive instructive comments, and that 5-star designs receive instruction at similar rates to 4- and 3-star designs, these comments are unlikely to be generating the results of this paper, which are large and increase dramatically from 3- to 5-stars. As a final check, in Table D.4 I expand the specifications in the previous table to match the format of the main results in the paper, interacting with competition and adding fixed effects and controls. The patterns in the previous table persist, and – importantly – we see no evidence that 5-star designs receive comments of any kind at different rates in the presence versus absence of top-rated competition.

Table D.4: Adding measures of competition, interactions, and controls

	Commented on	Instructive	Instructive
	(1)	(2)	(3)
Rated 5-stars	0.148***	0.100***	0.088
	(0.022)	(0.018)	(0.067)
* 1+ competing 5-stars	-0.013	-0.004	-0.203
	(0.027)	(0.020)	(0.273)
* prize value ($100s)	0.002	-0.004	-0.013
	(0.008)	(0.006)	(0.069)
Rated 4-stars	0.136***	0.108***	0.134***
	(0.015)	(0.011)	(0.047)
Rated 3-stars	0.092***	0.077***	0.133***
	(0.015)	(0.012)	(0.041)
Rated 2-stars	0.055***	0.047***	0.114**
	(0.016)	(0.012)	(0.046)
Rated 1-star	0.045**	0.020**	-0.066
	(0.020)	(0.010)	(0.059)
One or more competing 5-stars	0.014	0.008	-0.039
	(0.016)	(0.011)	(0.067)
Pct. of contest elapsed	-0.179***	-0.130***	-0.148**
	(0.014)	(0.011)	(0.061)
N	45170	45170	3537
R^2	0.26	0.20	0.45
Controls	Yes	Yes	Yes
Contest FEs	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes
Sample	All designs	All designs	Commented

Notes: Table provides results from regressing an indicator for whether a design received any comment (Column 1) or received a constructive comment (Columns 2 and 3) on indicators for its rating as well as an interaction for the presence of top-rated competition, as in the paper. Column (3) restricts the sample to only designs that received any comment. All columns include contest and player fixed effects and the standard set of controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by contest in parentheses.

E Robustness Checks (1)

The following tables provide robustness checks on the focal results in Section 3 estimating the effects of competition on similarity, using the difference hash algorithm. These estimates demonstrate that the results are not sensitive to the procedure used to calculate similarity scores. Table E.1 is a robustness check on Table 4; Table E.2, on Table 5; Table E.3, on Table 6; Table E.4, on Table 7; and Table E.5, on Table 8. The results in these tables are statistically and quantitatively similar to those in the body of the paper, despite the use of the computationally distinct similarity measure.

Table E.1: Similarity to player's best previously-rated designs (diff. hash)

	(1)	(2)	(3)	(4)	(5)
Player's prior best rating==5	0.329***	0.373***	0.201	0.246*	0.242*
	(0.117)	(0.116)	(0.124)	(0.133)	(0.141)
* 1+ competing 5-stars	-0.118	-0.152**	-0.126	-0.169**	-0.177**
	(0.084)	(0.074)	(0.079)	(0.086)	(0.087)
* prize value ($100s)	-0.021	-0.019	-0.006	-0.018	-0.024
	(0.027)	(0.028)	(0.036)	(0.038)	(0.042)
Player's prior best rating==4	0.168***	0.171***	0.095**	0.067*	0.049
	(0.036)	(0.035)	(0.039)	(0.039)	(0.041)
Player's prior best rating==3	0.124***	0.128***	0.069*	0.044	0.033
	(0.033)	(0.033)	(0.037)	(0.038)	(0.039)
Player's prior best rating==2	0.098***	0.104***	0.034	0.014	0.007
	(0.034)	(0.034)	(0.039)	(0.040)	(0.040)
One or more competing 5-stars	-0.018	-0.028	-0.003	-0.010	-0.019
	(0.020)	(0.028)	(0.024)	(0.031)	(0.032)
Prize value ($100s)	-0.023**		-0.016		
	(0.009)		(0.013)		
Pct. of contest elapsed	-0.055	-0.059*	-0.012	-0.021	-0.057
	(0.036)	(0.034)	(0.036)	(0.038)	(0.104)
Constant	0.496***	0.426***	0.506***	0.480***	0.491***
	(0.052)	(0.035)	(0.055)	(0.088)	(0.119)
N	3871	3871	3871	3871	3871
R^2	0.04	0.15	0.48	0.53	0.53
Contest FEs	No	Yes	No	Yes	Yes
Player FEs	No	No	Yes	Yes	Yes
Other Controls	No	No	No	No	Yes

Notes: Observations are designs. Dependent variable is a continuous measure of a design's similarity to the highest-rated preceding entry by the same player, taking values in [0,1], where a value of 1 indicates the design is identical to another. The mean value of this variable in the sample is 0.52 (s.d. 0.30). Column (5) controls for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a difference hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table E.2: Change in similarity to player's best previously-rated designs (diff. hash)

	(1)	(2)	(3)	(4)	(5)
Δ(Player's best rating==5)	0.635***	0.658***	0.682***	0.688**	0.693***
	(0.203)	(0.218)	(0.257)	(0.268)	(0.267)
* 1+ competing 5-stars	-0.352**	-0.349**	-0.374*	-0.364*	-0.368*
	(0.159)	(0.175)	(0.207)	(0.218)	(0.218)
* prize value ($100s)	-0.046	-0.048	-0.060	-0.062	-0.064
	(0.044)	(0.046)	(0.055)	(0.057)	(0.057)
Δ(Player's best rating==4)	0.244***	0.262***	0.235***	0.232***	0.232***
	(0.066)	(0.070)	(0.081)	(0.086)	(0.086)
Δ(Player's best rating==3)	0.175***	0.192***	0.168**	0.162**	0.162**
	(0.058)	(0.062)	(0.073)	(0.077)	(0.077)
Δ(Player's best rating==2)	0.121**	0.133**	0.109	0.104	0.104
	(0.053)	(0.058)	(0.067)	(0.071)	(0.071)
One or more competing 5-stars	-0.002	-0.002	0.004	-0.001	-0.001
	(0.008)	(0.017)	(0.017)	(0.028)	(0.029)
Prize value ($100s)	0.002		0.008		
	(0.003)		(0.009)		
Pct. of contest elapsed	-0.001	-0.007	-0.010	-0.014	-0.104
	(0.014)	(0.020)	(0.029)	(0.036)	(0.110)
Constant	-0.019*	-0.009	-0.033	0.031	0.116
	(0.011)	(0.010)	(0.032)	(0.089)	(0.123)
N	2694	2694	2694	2694	2694
R^2	0.01	0.04	0.10	0.13	0.13
Contest FEs	No	Yes	No	Yes	Yes
Player FEs	No	No	Yes	Yes	Yes
Other Controls	No	No	No	No	Yes

Notes: Observations are designs. Dependent variable is a continuous measure of the *change* in designs' similarity to the highest-rated preceding entry by the same player, taking values in [-1,1], where a value of 0 indicates that the player's current design is as similar to her best preceding design as was her previous design, and a value of 1 indicates that the player transitioned fully from innovating to recycling (and a value of -1, the converse). The mean value of this variable in the sample is -0.01 (s.d. 0.25). Column (5) controls for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a difference hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table E.3: Similarity to other designs in the same submission batch (diff. hash)

	Unweighted		Weighted	
	(1)	(2)	(3)	(4)
Player's prior best rating==5	0.225	0.246	0.235	0.260
	(0.299)	(0.293)	(0.286)	(0.281)
* 1+ competing 5-stars	-0.328**	-0.324**	-0.313**	-0.308**
	(0.146)	(0.144)	(0.147)	(0.145)
* prize value ($100s)	-0.022	-0.023	-0.025	-0.026
	(0.093)	(0.092)	(0.087)	(0.085)
Player's prior best rating==4	-0.015	-0.003	-0.013	0.004
	(0.031)	(0.032)	(0.030)	(0.031)
Player's prior best rating==3	0.011	0.019	0.010	0.020
	(0.034)	(0.035)	(0.031)	(0.032)
Player's prior best rating==2	-0.019	-0.012	-0.022	-0.014
	(0.047)	(0.049)	(0.045)	(0.045)
One or more competing 5-stars	-0.017	-0.018	-0.015	-0.017
	(0.033)	(0.034)	(0.032)	(0.033)
Pct. of contest elapsed	-0.001	-0.024	0.003	0.007
	(0.039)	(0.085)	(0.038)	(0.079)
Constant	0.643***	0.673***	0.670***	0.661***
	(0.121)	(0.156)	(0.099)	(0.128)
N	1987	1987	1987	1987
R^2	0.59	0.59	0.59	0.59
Contest FEs	Yes	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes	Yes
Other Controls	No	Yes	No	Yes

Notes: Observations are design batches, which are defined to be a set of designs by a single player entered into a contest in close proximity (15 minutes). Dependent variable is a continuous measure of intra-batch similarity, taking values in [0,1], where a value of 1 indicates that two designs in the batch are identical. The mean value of this variable in the sample is 0.69 (s.d. 0.28). Columns (3) and (4) weight the regressions by batch size. Columns (2) and (4) control for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a difference hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table E.4: Similarity to player's best not-yet-rated designs (placebo test; diff. hash)

	Similarity to forthcoming			Residual
	(1)	(2)	(3)	(4)
Player's best forthcoming rating==5	0.203	0.069	0.022	0.060
	(0.241)	(0.116)	(0.127)	(0.119)
* 1+ competing 5-stars	-0.040	-0.024	0.000	-0.023
	(0.145)	(0.073)	(0.080)	(0.079)
* prize value ($100s)	-0.064	-0.009	-0.001	-0.006
	(0.042)	(0.028)	(0.032)	(0.029)
Player's best forthcoming rating==4	0.023	0.051	0.059	0.045
	(0.074)	(0.066)	(0.064)	(0.067)
Player's best forthcoming rating==3	0.043	0.069	0.069	0.055
	(0.050)	(0.055)	(0.055)	(0.053)
Player's best forthcoming rating==2	0.031	0.025	0.026	0.025
	(0.048)	(0.051)	(0.051)	(0.049)
One or more competing 5-stars	-0.077	-0.089	-0.087	-0.085
	(0.076)	(0.115)	(0.119)	(0.124)
Pct. of contest elapsed	-0.210	0.033	0.093	0.023
	(0.261)	(0.442)	(0.422)	(0.469)
Constant	0.735***	0.448	0.423	-0.025
	(0.243)	(0.465)	(0.473)	(0.521)
N	1147	577	577	577
R^2	0.69	0.87	0.88	0.69
Contest FEs	Yes	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes	Yes

Table provides a placebo test of the effects of future feedback on similarity. Observations are designs. Dependent variable in Columns (1) to (3) is a continuous measure of a design's similarity to the best design that the player has previously entered that has yet to *but will eventually be* rated, taking values in [0,1], where a value of 1 indicates that the two designs are identical. The mean value of this variable is 0.50 (s.d. 0.29). Under the identifying assumption that future feedback is unpredictable, current choices should be unrelated to forthcoming ratings. Note that a given design's similarity to an earlier, unrated design can be incidental if they are both tweaks on a rated third design. To account for this possibility, Column (2) controls for the given and unrated designs' similarity to the best previously-rated design. Column (3) allows these controls to vary with the highest rating previously received. Dependent variable in Column (4) is the residual from a regression of the dependent variable in the previous columns on these controls. These residuals will be the subset of a given design's similarity to the unrated design that is not explained by jointly-occurring similarity to a third design. All columns control for days remaining and number of previous designs by the player and her competitors. Similarity scores in this table are calculated using a difference hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table E.5: Similarity to any of player's previous designs: 4-vs-4 (diff. hash)

	(1)	(2)	(3)
Player's prior best rating==4	0.087	0.221*	0.745***
	(0.068)	(0.122)	(0.253)
* 1+ competing 4- or 5-stars	-0.105**	-0.190**	-0.304
	(0.047)	(0.084)	(0.201)
* prize value ($100s)	0.006	-0.007	-0.162*
	(0.020)	(0.037)	(0.084)
Player's prior best rating==3	-0.014	-0.009	-0.048
	(0.023)	(0.039)	(0.070)
Player's prior best rating==2	-0.016	0.049	0.058
	(0.026)	(0.053)	(0.097)
One or more competing 4- or 5-stars	0.045*	0.109	0.111
	(0.025)	(0.081)	(0.177)
Pct. of contest elapsed	-0.006	-0.394	1.045
	(0.101)	(0.267)	(0.864)
Constant	0.517***	1.033***	-0.621
	(0.168)	(0.294)	(0.717)
N	2926	1557	879
R^2	0.55	0.61	0.70
Contest FEs	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes
Other Controls	Yes	Yes	Yes
Restriction	All	2nd half	4th qtr

Notes: Table shows the effects of 4-star feedback and competition on similarity when no player has a 5-star rating. Observations are designs. Dependent variable is a continuous measure of a design's maximal similarity to previous entries in the same contest by the same player, taking values in [0,1], where a value of 1 indicates the design is identical to another. All columns include contest and player fixed effects and control for the number of days remaining and number of previous designs entered by the player and her competitors. Columns (2) and (3) restrict the sample to submissions in the second half or fourth quarter of a contest, when the absence of 5-star ratings may be more meaningful and is increasingly likely to be final. Similarity scores in this table are calculated using a difference hash algorithm. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

F Robustness Checks (2)

The following tables show that the full effect of competition on high performers' originality is realized with one high-quality competitor, and does not vary as competition intensifies. Tables F.1 to F.3 demonstrate this result with the perceptual hash similarity measures. In all cases, I estimate differential patterns in the presence of one vs. two or more top-rated, competing designs and find no such difference.

Table F.1: Similarity to player's best previously-rated designs (p. hash)

	(1)	(2)	(3)	(4)	(5)
Player's prior best rating==5	0.442***	0.461***	0.259***	0.358***	0.362***
	(0.103)	(0.093)	(0.098)	(0.098)	(0.103)
* 1+ competing 5-stars	-0.252***	-0.290***	-0.176*	-0.225**	-0.226**
	(0.093)	(0.078)	(0.090)	(0.093)	(0.093)
* 2+ competing 5-stars	0.066	0.056	0.024	0.025	0.023
	(0.075)	(0.078)	(0.097)	(0.094)	(0.092)
* prize value ($100s)	-0.025	-0.016	0.005	-0.014	-0.018
	(0.025)	(0.024)	(0.032)	(0.031)	(0.033)
Player's prior best rating==4	0.165***	0.160***	0.127***	0.122***	0.116***
	(0.024)	(0.022)	(0.030)	(0.031)	(0.032)
Player's prior best rating==3	0.080***	0.077***	0.068**	0.061**	0.056**
	(0.018)	(0.019)	(0.028)	(0.028)	(0.028)
Player's prior best rating==2	0.045**	0.044**	0.022	0.026	0.024
	(0.021)	(0.022)	(0.029)	(0.030)	(0.030)
One or more competing 5-stars	-0.025	0.017	0.005	0.002	0.002
	(0.030)	(0.033)	(0.032)	(0.037)	(0.037)
Two or more competing 5-stars	0.008	-0.012	-0.011	0.004	0.000
	(0.030)	(0.036)	(0.034)	(0.042)	(0.044)
Prize value ($100s)	-0.014*		-0.010		
	(0.007)		(0.010)		
Pct. of contest elapsed	-0.031	-0.059*	-0.010	-0.019	-0.103
	(0.034)	(0.033)	(0.030)	(0.034)	(0.084)
Constant	0.238***	0.207***	0.234***	0.231***	0.301***
	(0.039)	(0.024)	(0.044)	(0.062)	(0.094)
N	3871	3871	3871	3871	3871
R^2	0.07	0.20	0.48	0.53	0.53
Contest FEs	No	Yes	No	Yes	Yes
Player FEs	No	No	Yes	Yes	Yes
Other Controls	No	No	No	No	Yes

Notes: Observations are designs. Dependent variable is a continuous measure of a design's similarity to the highest-rated preceding entry by the same player, taking values in [0,1], where a value of 1 indicates the design is identical to another. The mean value of this variable in the sample is 0.28 (s.d. 0.27). Column (5) controls for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table F.2: Change in similarity to player's best previously-rated designs (p. hash)

	(1)	(2)	(3)	(4)	(5)
Δ(Player's best rating==5)	0.860***	0.879***	0.936***	0.921***	0.928***
	(0.157)	(0.166)	(0.194)	(0.200)	(0.200)
* 1+ competing 5-stars	-0.497***	-0.487***	-0.535***	-0.504***	-0.505***
	(0.119)	(0.133)	(0.150)	(0.160)	(0.161)
* 2+ competing 5-stars	0.106	0.099	0.163	0.108	0.107
	(0.106)	(0.111)	(0.123)	(0.127)	(0.127)
* prize value ($100s)	-0.090**	-0.093**	-0.115***	-0.108**	-0.110**
	(0.036)	(0.037)	(0.044)	(0.044)	(0.045)
Δ(Player's best rating==4)	0.279***	0.285***	0.274***	0.281***	0.284***
	(0.061)	(0.065)	(0.071)	(0.077)	(0.077)
Δ(Player's best rating==3)	0.146***	0.154***	0.140**	0.141**	0.142**
	(0.054)	(0.057)	(0.063)	(0.067)	(0.067)
Δ(Player's best rating==2)	0.082*	0.085*	0.069	0.064	0.063
	(0.042)	(0.046)	(0.052)	(0.056)	(0.056)
One or more competing 5-stars	-0.004	0.002	-0.017	-0.014	-0.013
	(0.016)	(0.026)	(0.031)	(0.044)	(0.043)
Two or more competing 5-stars	0.001	-0.006	0.020	0.031	0.029
	(0.018)	(0.025)	(0.033)	(0.041)	(0.043)
Prize value ($100s)	0.003		0.003		
	(0.003)		(0.008)		
Pct. of contest elapsed	0.015	0.009	0.014	-0.003	-0.050
	(0.012)	(0.018)	(0.024)	(0.029)	(0.074)
Constant	-0.029***	-0.017*	-0.030	0.062	0.104
	(0.011)	(0.010)	(0.029)	(0.092)	(0.108)
N	2694	2694	2694	2694	2694
R^2	0.03	0.05	0.11	0.14	0.14
Contest FEs	No	Yes	No	Yes	Yes
Player FEs	No	No	Yes	Yes	Yes
Other Controls	No	No	No	No	Yes

Notes: Observations are designs. Dependent variable is a continuous measure of the *change* in designs' similarity to the highest-rated preceding entry by the same player, taking values in [-1,1], where a value of 0 indicates that the player's current design is as similar to her best preceding design as was her previous design, and a value of 1 indicates that the player transitioned fully from innovating to recycling (and a value of -1, the converse). The mean value of this variable in the sample is 0.00 (s.d. 0.23). Column (5) controls for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table F.3: Similarity to other designs in the same submission batch (p. hash)

	Unweighted		Weighted	
	(1)	(2)	(3)	(4)
Player's prior best rating==5	0.213	0.240	0.245	0.285
	(0.312)	(0.307)	(0.302)	(0.294)
* 1+ competing 5-stars	-0.437**	-0.433**	-0.475**	-0.469**
	(0.217)	(0.214)	(0.203)	(0.200)
* 2+ competing 5-stars	0.180	0.177	0.237	0.238
	(0.236)	(0.232)	(0.223)	(0.218)
* prize value ($100s)	0.020	0.015	0.013	0.009
	(0.099)	(0.099)	(0.095)	(0.093)
Player's prior best rating==4	0.055*	0.066*	0.065**	0.086**
	(0.032)	(0.037)	(0.032)	(0.038)
Player's prior best rating==3	0.056	0.062*	0.052	0.065*
	(0.035)	(0.037)	(0.035)	(0.037)
Player's prior best rating==2	0.023	0.029	0.009	0.020
	(0.049)	(0.050)	(0.047)	(0.047)
One or more competing 5-stars	0.083	0.086	0.076	0.078
	(0.063)	(0.063)	(0.063)	(0.063)
Two or more competing 5-stars	-0.094	-0.100	-0.080	-0.086
	(0.072)	(0.076)	(0.074)	(0.076)
Pct. of contest elapsed	-0.016	-0.102	-0.005	-0.064
	(0.049)	(0.115)	(0.050)	(0.112)
Constant	0.392***	0.504***	0.385***	0.457***
	(0.066)	(0.149)	(0.061)	(0.148)
N	1987	1987	1987	1987
R^2	0.58	0.58	0.58	0.58
Contest FEs	Yes	Yes	Yes	Yes
Player FEs	Yes	Yes	Yes	Yes
Other Controls	No	Yes	No	Yes

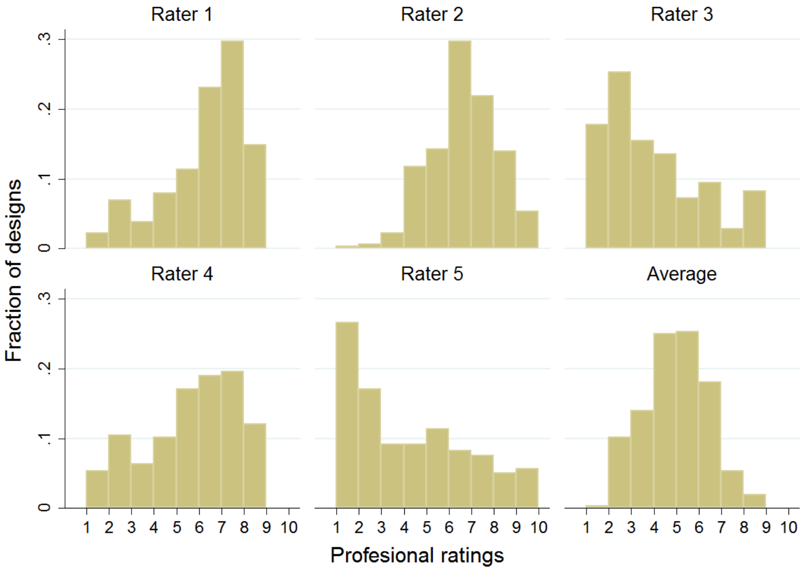
Notes: Observations are design batches, which are defined to be a set of designs by a single player entered into a contest in close proximity (15 minutes). Dependent variable is a continuous measure of intra-batch similarity, taking values in [0,1], where a value of 1 indicates that two designs in the batch are identical. The mean value of this variable in the sample is 0.45 (s.d. 0.32). Columns (3) and (4) weight the regressions by batch size. Columns (2) and (4) control for the number of days remaining and number of previous designs entered by the player and her competitors. Similarity scores in this table are calculated using a perceptual hash algorithm. Preceding designs/ratings are defined to be those entered/provided at least 60 minutes prior to the given design. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

G Collection of Professional Ratings

The panelists participating in the ratings exercise were recruited through the author's personal and professional networks and hired at their regular rates. All have formal training and experience in graphic design, and they represent a diverse swath of the profession: three panelists work at advertising agencies, and two others are employed in-house for a client and primarily as a freelancer (respectively).

Ratings were collected though a web-based application. Designs were presented to each panelist in random order, and panelists were limited to 100 ratings per day. With each design, the panelist was provided the project title and client industry (as they appeared in the design brief in the source data) and instructed to rate the "quality and appropriateness" of the given logo on a scale of 1 to 10. Panelists were asked to rate each logo "objectively, on its own merits" and not to "rate logos relative to others." Figure G.1 provides the distribution of ratings from each of the five panelists and the average.

Figure G.1: Panelists' ratings on subsample of sponsors' top-rated designs



Notes: Figure shows the distribution of professionals' ratings on all 316 designs in the dataset that received the top rating from contest sponsors. Professional graphic designers were hired at regular rates to participate in this task. Each professional designer provided independent ratings on every design in the sample rated 5 stars by a contest sponsor. Ratings were solicited on a scale of 1-10, in random order, with a limit of 100 ratings per day.

It can be seen in the figure that one panelist ("Rater 5") amassed over a quarter of her ratings at the lower bound, raising questions about the reliability of these assessments: it is unclear whether the panelist would have chosen an even lower rating had the option been available. The panelist's tendency to assign very low ratings became apparent after the first day of her participation, and in light of the anomaly, the decision to omit this panelist's ratings from the analysis was made at that time. The paper's results are nevertheless robust to including ratings from this panelist that lie above the lower bound.