**Essays in Innovation, Past and Present**

by

Daniel Pincus Gross

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Benjamin R. Handel, Chair
Professor Barry J. Eichengreen
Professor John Morgan
Professor Steven Tadelis

Spring 2015

**Essays in Innovation, Past and Present**

**Abstract**

Essays in Innovation, Past and Present

by

Daniel Pincus Gross

Doctor of Philosophy in Economics

University of California, Berkeley

Professor Benjamin R. Handel, Chair

This dissertation studies the economics of historical and modern innovation. The first chapter makes inroads into understanding how competition and incentives shape the creative process which lies at the heart of all technological progress. The creative act is a classic example of a black box in academic research: we can see the inputs and outputs, but we know little about what happens in between. This paper uses new tools for measuring the content of digital media to see how commercial graphic designers' work evolves in winner-take-all competition. In this chapter, I show that competition both creates and destroys incentives for innovation: some competition is necessary to motivate high-performers to experiment with novel, untested ideas over tweaking tried-and-true approaches, but heavy competition will drive them out of the market.

In the second chapter, I study the effects of performance feedback on innovation in competitive settings. Feedback typically serves two functions: it informs agents of their relative performance, and it also helps them improve the quality of their product. The presence of these effects suggests a tradeoff between participation and improvement, as the revelation of asymmetries discourages effort. Using data from the same setting as chapter one, I first show that this tradeoff is real. I then develop a structural model of the setting – the first of its kind in the literature – and use the results to evaluate counterfactual feedback policies. The results suggest that feedback is on net a desirable mechanism for a principal seeking high-quality innovation.

In the third chapter, I use the farm tractor as a case study to demonstrate that technologies diffuse along two distinct margins: scale and scope. Although tractors are now used in nearly every field operation and with nearly all crops, early models were far more limited in their capabilities, and only in the late 1920s did the technology begin to generalize for broader use with row crops such as corn. Diffusion prior to 1930 was accordingly heavily concentrated in the Wheat Belt, while growth in diffusion from 1930-1940 was concentrated in the Corn Belt. Other historically important innovations in agriculture and manufacturing share similar histories of expanding scope. The key to understanding the pace and path of technology diffusion is thus not only in explaining the number of different users, but also in explaining the number of different *uses*.

A common theme across all three chapters is the focus on developing tools or strategies to study innovation that are less dependent on patent data than the extant literature, since the majority of innovation is not patented (and often not patentable), and doing so while advancing the empirical literature on innovation in new directions.

To my mom, for making this possible. I love you.

## Acknowledgments

I am indebted to Ben Handel and Barry Eichengreen for their wisdom, guidance, and support over the last five years, in all the forms it has taken. To John Morgan, whose conversation and feedback provided much of the inspiration my job market paper, and whose encouragement gave me – and allowed me to give myself – the freedom to think outside the box. To Steve Tadelis, for his feedback and advice in shepherding me through the job market. And to Marty Olney, for her tireless devotion to Berkeley graduate students – from training in pedagogy to economic history discussion groups and lunches. I also thank Patrick Allen and Vicky Lee for the incredible work they do for the economics graduate students. I am grateful to the many colleagues who have shared feedback on the chapters of this dissertation, including Dominick Bartelme, Youssef Benzarti, Carola Binder, Ernesto Dal Bó, Bo Cowgill, Jose Espin-Sanchez, Joe Farrell, Przemek Jeziorski, Gustavo Manso, Denis Nekipelov, Matthew Rabin, Gautam Rao, Reed Walker, Glen Weyl, Brian Wright, Noam Yuchtman, participants at the numerous venues where I have had the opportunity to present the work herein, and many others. The third chapter has especially benefited from the comments of Susan Carter, Brad DeLong, Alex Field, Joel Mokyr, Petra Moser, Alan Olmstead, Paul Rhode, Daniel Robert, and Richard Sutch. I owe a special thanks to my lifelong friend Jared Scheib, for challenging me to question everything and pushing the limits of my intellect, communication skills, and patience. This research was supported by UC Berkeley, the National Science Foundation, the Economic History Association, the All-UC Group in Economic History, and the Berkeley Economic History Lab. As they say: all errors are mine and mine alone.

*Veni, vedi, vici.*

# Part I

# Creativity Under Fire: The Effects of Competition on the Creative Process

# Chapter Abstract

Creativity is fundamental to innovation and pervasive in everyday life, yet the creative process has received only limited attention in economics and can in practice be difficult to model and measure. In this paper, I study the effect of competition on individuals' incentives for creative experimentation in the production of commercial art. Using a sample of logo design contests, and a novel, content-based measure of designs' originality, I find that competition has an inverted-U shaped effect on individuals' propensity for innovation: some competition is necessary to induce players to experiment with novel, untested ideas, but heavy competition can drive them to abandon the tournament altogether, such that experimentation is maximized by the presence of one high-quality competitor. The evidence is consistent with a generalized model of agents' choice between risky, radical innovation; more reliable, incremental innovation; and exit from a creative tournament where agents are risk-averse or face decreasing returns to improvement due to a concave success function. These results reconcile conflicting evidence from an extensive literature on the effects of competition on innovation and have direct implications for R&D policy, competition policy, and the management of organizations in creative or research industries.

The creative act is among the most important yet least understood phenomena in economics and the social and cognitive sciences. Technological progress – the wellspring of lasting, long-run economic growth – at heart consists of creative solutions to familiar problems. Millions of people in the U.S. alone work in creative fields ranging from research, to software development, to media and the arts,[1] and surveys show CEOs' top concerns consistently include creativity and innovation in the firm.[2] Despite its importance, the creative process has received only limited attention in economic research and has historically proven difficult to both model and measure.

In this paper, I study the incentive effects of competition on creative production. I do so in an empirical setting where creative experimentation and competition can be both precisely measured and disentangled: commercial graphic design tournaments. Using image comparison algorithms to measure experimentation, I provide causal evidence that competition can both create and destroy incentives for innovation. I find that some competition is necessary for high-performing agents to prefer experimenting with novel, untested ideas over tweaking their earlier work, but that heavy competition discourages effort of any kind. These patterns are driven by risk-return tradeoffs inherent to innovation, which I show to be high-risk, high-return. The implication of these results is an inverted-U shaped effect of competition on innovation, with incentives for taking creative risks maximized by the presence of *one* high-quality competitor.

The challenge of motivating creativity can be naturally characterized as a principal-agent problem. Suppose a firm wants its workers to experiment with new, potentially better (lower cost or higher quality) product designs, but the firm does not observe workers' creative choices and can only reward them on the quality of their output. In this setting, failed experimentation is indistinguishable from shirking. Workers who are risk-averse or face decreasing returns to improvement, as they do in this paper, may then prefer exploiting existing solutions over experimenting if the existing method reliably yields an acceptable result – even if creative and routine effort are equally expensive. Motivating innovation will be even more difficult when creative effort is more costly.

To better understand the economics of the creative process, I begin by developing a model of a winner-take-all "creative tournament," building on the economics literature on innovation and tournament competition.[3] In the model, a principal seeks a high-value product design from a pool of workers and solicits ideas using a fixed-length tournament mechanism, awarding a prize to the preferred entry. Workers compete for the prize by entering designs in turns. At each turn, a worker must choose between experimenting with a new design, tweaking an existing design, or

[1] According to the U.S. Census 2010 County Business Patterns, over 15 million people are employed in the Media and Communications; Professional, Scientific, and Technical Services; Management; and Arts and Entertainment sectors alone – fields that could be considered creative professions. This total represents nearly 15% of U.S. employment and over 20% of wages, and it excludes industries in which creativity may be valued but is not strictly essential.

[2] Annual CEO surveys by The Conference Board reveal "stimulating innovation/creativity/enabling entrepreneurship" is consistently among executives' top concerns. "Innovation" was perceived as the top global challenge in 2012 and the third biggest challenge in 2013 and 2014. See http://www.conference-board.org/subsites/index.cfm?id=14514.

[3] Since the seminal contributions of Arrow (1962), countless papers have studied incentives for innovation. Wright (1983) presents an interesting theoretical comparison of patents, prizes, and research contracts as incentive mechanisms, and Scotchmer (2004) provides a summary of the literature. The model in this paper is most closely related to the work of Taylor (1995), Che and Gale (2003), Fullerton and McAfee (1999), and Terwiesch and Xu (2008). Though this mechanism is commonly described in the economics literature as an "R&D," "research," or "innovation" tournament, I refer to it in the remainder of this paper as a "creative" tournament to emphasize that it applies to creative production of all kinds – not strictly research or product development. The model in this paper also has ties to recent work on tournaments with feedback, such as Yildirim (2005) and Ederer (2010), where agents accumulate effort over multiple rounds with interim evaluation.

abandoning the tournament altogether.[4] Each submission receives immediate, public feedback on its quality, and at the end of the contest, the sponsor selects the winner. The model establishes a new explanation for an inverted-U relationship between competition and innovation, with incentives for experimentation maximized at intermediate levels of competition.[5]

I then bring the theoretical intuition to an empirical study of design competitions similar to the model's setting, using a sample of logo design contests from a prominent online platform.[6] In these contests, a firm ("sponsor") solicits custom designs from a community of freelance designers ("players") in exchange for a winner-take-all prize. The contests in the sample offer prizes of a few hundred dollars and on average attract nearly 35 players and 100 designs. An important feature of this setting is that the sponsor can provide interim 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 level of competition they face. The dataset also includes the designs themselves, allowing me to study experimentation in this venue: 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 in a contest relative to prior designs by that player and her competitors.

This setting presents a unique opportunity to directly observe creative experimentation in the field. Though production of commercial advertising is interesting in its own right – advertising is a $120 billion industry in the U.S. and a $520 billion industry worldwide[7] – the design process observed here is similar to that in other settings where new products are developed. It also has parallels to the experimentation with inputs and production techniques responsible for productivity improvements in firms, including those not strictly in the business of producing cutting-edge ideas: Hendel and Spiegel (2014) study plant-level productivity at a steel mill and suggest that a large fraction of its unexplained TFP growth results from the accumulation of adjustments to its production process that are tested, evaluated, and implemented over time.

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: a five-star design has 10 times the weight of a four-star design, 100 times that of a three-star design, and nearly 2,000 times that of a one-star design in the success function. Data on the time at which designs are entered by players and rated by sponsors enables me to determine what every participant knows at each point in time – and what they have yet to find out. To obtain causal

[4]The explore-exploit dilemma is endemic to a class of decision models known as bandit problems, which have received extensive coverage in the economics, statistics, and operations research literatures. Weitzman (1979) provides one of the earliest applications in economics, examining optimal stopping rules in a sequential search for an innovation. Manso (2011) and Ederer and Manso (2013) study incentives for exploration in a single-agent, dynamic two-armed bandit model and an accompanying lab experiment and find that the optimal contract for motivating innovation tolerates early failure and rewards long-term success. See Bergemann and Valimaki (2008) for other examples.

[5]These results concord with the standard result from the tournament literature that asymmetries discourage effort (e.g., Baik 1994, Brown 2011). The contribution of this paper is to embed an explore-exploit problem in the model, effectively adding a new margin along which effort may vary: radical versus incremental. I show that intermediate competition not only maximizes incentives to participate; it also maximizes incentives to *experiment*.

[6]The empirical setting is conceptually similar to the computer programming competitions studied by Boudreau et al. (2011), Boudreau et al. (2014), and Boudreau and Lakhani (2014). Logo design competitions have also recently been studied in the management literature as examples of innovation tournaments (Wooten and Ulrich 2014).

[7]Magna Global Advertising Forecast for 2014, available at http://news.magnaglobal.com/.

estimates of the effects of feedback and competition, I exploit naturally-occurring, quasi-random variation in the timing of sponsors' ratings, and identify the effects of ratings they can observe. The empirical strategy effectively compares players' responses to feedback and competition they observe at the time of design against that which has not yet been given.

I find that feedback and competition have significant effects on creative choices. In the absence of competition, positive feedback causes players to cut back sharply on experimentation: players with the top rating enter designs that are one full standard deviation more similar to their previous work than those who have 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 one, on average *three* standard deviations more similar to it – and attenuates at each rung down the ratings ladder. But these effects are significantly reversed (by half or more) when high-quality competition is present. Intense competition and negative feedback also drive players to stop investing, as their work is unlikely to win; the probability of abandonment increases with each high-rated competitor. In both reduced-form regressions and a descriptive choice model, I find high-performers are most likely to experiment when they face exactly one high-quality competitor.

For players with poor designs, the data show that continued experimentation clearly dominates imitation of the poor-performing work. But why would a top contender ever deviate from her winning formula? The model suggests that even a top contender may wish to experiment when competition is present, *provided there is sufficient upside to experimentation.* To evaluate whether it pays to innovate, I recruit a panel of professional designers to provide independent ratings on all five-star designs in my sample and correlate their responses with these designs' originality. I find that experimentation on average results in higher-rated designs than incremental changes but that the distribution of opinion also has higher variance. These results validate one of the standard assumptions in the innovation literature – that experimentation is high-risk and high-reward – which is the necessary condition for competition to motivate innovation.

To my knowledge, this paper provides the most direct view into the creative process to-date in the economics literature. The creative act is a classic example of a black box: we can see the inputs and outputs, but we have little evidence or understanding of what happens in between. Reflecting these data constraints, empirical research has opted to measure innovation in terms of inputs (R&D spending) and outcomes (patents), when innovation is at heart about what goes on in between: individual acts of discovery and invention. Because experimentation choices cannot be inferred from R&D inputs alone, and because patent data only reveal the successes – and only the subset that are patentable and its owners are willing to disclose – we may know far less about innovation than commonly believed. This paper is an effort to fill this gap.

While creativity has only recently begun to receive attention from economists,[8] social psychologists have studied the effects of intrinsic and extrinsic motivation on creativity for decades.[9] The consensus from this literature is that creativity is inspired by intrinsic "enjoyment, interest, and personal challenge" (Hennessey and Amabile 2010), and that extrinsic pressures of reward, supervision,

[8]For examples, see Weitzman (1998) and Azoulay et al. (2011). Akcigit and Liu (2014) and Halac et al. (2014) study problems more similar to the one in this paper: the former embed an explore-exploit problem into a two-player patent race, as risky and safe lines of research, and study the efficiency consequences of private information; the latter study the effects of various disclosure and prize-sharing policies on effort in contests for innovation. Charness and Grieco (2014) find that financial incentives can elicit "closed" (targeted) creativity but not "open" (blue-sky) thinking. Mokyr (1990) provides a history of technological creativity at the societal level, dating to classical antiquity.

[9]See Hennessey and Amabile (2010) for a comprehensive review of creativity research in psychology.

evaluation, and competition tend to undermine intrinsic motivation by causing workers to "feel controlled by the situation." The implication is that creativity cannot be managed: any attempts to manage creativity will backfire, and the best one can do is to provide a supportive environment for creative workers, leave them alone, and hope for the best.[10] Although intrinsic motivation is undoubtedly important to creativity, I counter these claims with evidence that individuals' creative choices respond positively to well-designed incentive schemes.[11]

The evidence that incentives for assuming creative risk are highest with moderate competition has broader implications for R&D policy, competition policy, and management of organizations in creative and research industries, which I discuss in Section I.6. The results also provide a partial resolution to the long-standing debate on the effects of competition on innovation, summarized by Gilbert (2006) and Cohen (2010). Since Schumpeter's contention that monopoly is most favorable to innovation, researchers have produced explanations for and empirical evidence of positive, negative, and inverted-U relationships between competition and innovation. The confusion results from disagreements of definition and measurement; ambiguity in the type of competition being studied; problems with econometric identification; and institutional differences, such as whether innovation is appropriable. This paper addresses these issues by establishing clear and precise measures of competition and innovation, identifying the causal effects of *information* about competition on innovation, and focusing the analysis on a setting with a fixed, winner-take-all prize and copyright protections. Moreover, as Gilbert (2006) notes, the literature has largely ignored that individuals are the source of innovation ("discoveries come from creative people"), even if patents get filed by corporations. It is precisely this gap that I seek to fill with the present paper.

The paper proceeds as follows. Section I.1 presents the model of winner-take-all creative competition. Section I.2 introduces the empirical setting, including my approach to measuring experimentation, and describes the identification strategy. Section I.3 estimates the effects of competition on creative experimentation and participation. Section I.4 establishes that experimentation in this setting is high-risk, high-return, confirming the driving assumption of the model. In Section I.5, I unify these results and show that experimentation is maximized with one high-quality competitor. Section I.6 discusses implications of these results for policymakers, managers, and future research on innovation and the creative process. Section I.7 concludes with several questions on the creative act that I believe are ripe for attention.

## I.1 A Model of a Creative Tournament

Suppose a risk-neutral principal seeks to develop a new product design. Because R&D is risky, and designs are difficult to objectively value, the principal cannot contract directly on performance. It instead sponsors a tournament to solicit prototypes from a pool of $J$ risk-neutral players, who enter designs in turns and receive immediate, public feedback on their quality (defined below). Each

[10] As Amabile and Khaire (2008) write, "One doesn't manage creativity. One manages for creativity."

[11] The findings of this paper are not unprecedented. A smaller, rival camp of psychologists has argued that reward can have profound effects on creativity, if only applied the right way: Eisenberger and Rhoades (2001) show in a series of experiments that creativity is enhanced by rewards when it is clear to participants that creative performance is precisely what is being rewarded. This is not to say that intrinsic motivators are unimportant or should be disregarded (e.g., as Stern (2004) shows, corporate scientists sacrifice wages for the opportunity to conduct self-directed research and publish) but rather that even creative types appreciate, and will compete for, rewards of money, status, and recognition, be it out of self-interest, a desire to share the value of one's discovery or creation with a broader audience, or both.

design in the competition is either generated by experimentation, which has stochastic outcomes, or incrementally adapted from the blueprints of previous entries; players who choose to continue working on a given design post-feedback can re-use the blueprint to create variants, though the original version remains in contention. At a given turn, the player must choose whether to continue participating and if so, what type of innovation to undertake: radical or incremental. At the end of the tournament, the sponsor awards a fixed, winner-take-all prize $P$ to its favorite entry. The sponsor seeks to maximize the value of the winning design.

To hone intuition, suppose each player enters at most two designs. Let each design be characterized by latent value $\nu_{jt}$, which only the sponsor observes (possibly sponsor-specific):

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

where $j$ indexes players and $t$ indexes designs. In this model, $\beta_{jt}$ represents the design's quality, which may not be known ex-ante and is revealed by the sponsor's feedback. The design's value to the sponsor, $\nu_{jt}$, is increasing and concave in its quality, and the design with the highest $\nu$ wins the contest.[12] The $\varepsilon_{jt}$ term is a random shock, which can be interpreted as idiosyncracies in the sponsor's tastes at the time a winner is chosen. Player $j$'s probability of winning is then:

$$Pr\,(\text{player } j \text{ wins}) = \frac{\beta_{j1} + \beta_{j2}}{\sum_{k \neq j} (\beta_{k1} + \beta_{k2}) + \beta_{j1} + \beta_{j2}} = \frac{\beta_{j1} + \beta_{j2}}{\mu_j + \beta_{j1} + \beta_{j2}} \tag{I-2}$$

where $\mu_j \equiv \sum_{k \neq j} (\beta_{k1} + \beta_{k2})$ is the competition that player $j$ faces in the contest.[13] This function is concave in the player's own quality and decreasing in the quality of competition.

Players develop and submit designs one at a time, in turns, and immediately receive public feedback that reveals $\beta_{jt}$. It is assumed that property protections are in place to prevent idea theft by competitors. Every player's first design in the contest is therefore novel to that contest, and at their subsequent turn, players have three options: they can exploit (tweak, or adapt) the existing design, explore (experiment with) an entirely new design, or abandon the contest altogether. I elaborate on each of these options below.

1. **E**xploitation is undertaken at cost $c > 0$ and yields a design concept of the same quality as the one being exploited. A player who chooses to exploit will tweak her first design, which has quality $\beta_{j1}$, resulting in a second-round design with $\beta_{j2} = \beta_{j1}$ and a new draw of the luck term, and increasing her probability of winning accordingly.

   After exploitation, the player's expected probability of winning is:

$$E\left[Pr\,(\text{player } j \text{ wins} \mid \text{exploit})\right] = \frac{\beta_{j1} + \beta_{j1}}{\mu_j + \beta_{j1} + \beta_{j1}} \tag{I-3}$$

[12] The decision to model designs' latent value ($\nu_{jt}$) as a function of logged quality ($\beta_{jt}$) is taken for analytical convenience but also supported by the intuition of decreasing returns to quality. $\nu_{jt}$ could also be linear in $\beta_{jt}$ and similar results would obtain: the feature of the model driving the results is the concavity of the success function.

[13] Note that the level of competition is determined by both the number and quality of competing designs. As Section I.2 shows, a single, high-quality design can present an equal amount of competition as several lower-quality ones.

2. **E**xploration costs $d \geq c$ and yields either a high- or low-quality design concept, each with positive probability. Define $\alpha \geq 1$ as the exogenous degree of experimentation under this option (conversely, $\frac{1}{\alpha} \in [0,1]$ can be interpreted as the similarity of the new, experimental design and the player's first design). With probability $q$, experimentation will yield a high-quality design with $\beta_{j2}^{H} = \alpha\beta_{j1}$, and with probability $(1-q)$ it will yield a low-quality design with $\beta_{j2}^{L} = \frac{1}{\alpha}\beta_{j1}$. I assume $q > \frac{1}{1+\alpha}$, which implies that a design's expected quality under exploration ($E[\beta_{j2}|\text{Explore}]$) is greater than that under exploitation ($\beta_{j1}$). A risk-neutral sponsor will therefore *always* want players to explore. Note that as written, exploitation is a special case of exploration, with $\alpha = 1$.[14]

   After exploration, the player's expected probability of winning is:

$$E\left[Pr\left(\text{player } j \text{ wins} \mid \text{explore}\right)\right] = q \cdot \left(\frac{\beta_{j1} + \beta_{j2}^{H}}{\mu_j + \beta_{j1} + \beta_{j2}^{H}}\right) + (1-q) \cdot \left(\frac{\beta_{j1} + \beta_{j2}^{L}}{\mu_j + \beta_{j1} + \beta_{j2}^{L}}\right) \quad \text{(I-4)}$$

3. **A**bandonment is costless: the player can always walk away. Doing so leaves the player's probability of winning unchanged, as her earlier work remains in contention.

   After abandonment, the player's probability of winning will be:

$$E\left[Pr\left(\text{player } j \text{ wins} \mid \text{abandon}\right)\right] = \frac{\beta_{j1}}{\mu_j + \beta_{j1}} \quad \text{(I-5)}$$

In this setting, feedback has three effects: it informs each player about her first design's quality, helps her improve and set expectations over her second design, and reveals the level of competition. 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. The model thus characterizes incentives for innovation.

In the remainder of this section, I examine a player's incentives to explore, exploit, or abandon the competition. Section I.1 studies the conditions required for the player to prefer exploration over the alternatives and shows that these conditions lead to an inverted-U relationship between competition and innovation (proofs in Appendix I.B). Section I.1 contextualizes this result in the existing literature. To simplify the mathematics, I assume the level of competition $\mu_j$ is known to player $j$, though the results are general to other assumptions about players' beliefs over the competition they will face, including competitors' best responses. The model can also be extended to allow players to enter an arbitrary number of designs, and the results will hold as long as players do not choose exploration for its option value.

### Incentives for Radical Innovation

To simplify notation, let $F(\beta_2) = F(\beta_2|\beta_1, \mu)$ denote player $j$'s probability of winning with a second design of quality $\beta_2$, given $\beta_1$ and $\mu$ (omitting the $j$ subscript). The model permits four values of $\beta_2$: $\beta_2^H$, $\beta_2^L$, $\beta_1$, and 0. The first two values result from exploration, and the latter two from

[14] In this model, I assume $\alpha$ and $q$ are fixed. When $\alpha$ is endogenized and costless, the (risk-neutral) player's optimal $\alpha$ is infinite, since the experimental upside would then be unlimited and the downside bounded at zero. A natural extension would be to relax experimentation costs $d(\cdot)$ and/or the probability of a successful experiment $q(\cdot)$ to vary with $\alpha$. Such a model is considerably more difficult to solve and beyond the scope of this paper.

exploitation and abandonment, respectively. For player $j$ to experiment on a new design, she must prefer exploration over both exploitation (incentive compatibility) and abandonment (individual rationality):

$$\underbrace{\left[qF\left(\beta_2^H\right)+(1-q)\,F\left(\beta_2^L\right)\right]\cdot P-d}_{E[\pi|\text{explore}]}>\underbrace{F\left(\beta_1\right)\cdot P-c}_{E[\pi|\text{exploit}]} \tag{IC}$$

$$\underbrace{\left[qF\left(\beta_2^H\right)+(1-q)\,F\left(\beta_2^L\right)\right]\cdot P-d}_{E[\pi|\text{explore}]}>\underbrace{F\left(0\right)\cdot P}_{E[\pi|\text{abandon}]} \tag{IR}$$

These conditions can be rearranged to be written as follows:

$$qF\left(\beta_2^H\right)+(1-q)\,F\left(\beta_2^L\right)-F\left(\beta_1\right)>\frac{d-c}{P} \tag{IC}$$

$$qF\left(\beta_2^H\right)+(1-q)\,F\left(\beta_2^L\right)-F\left(0\right)>\frac{d}{P} \tag{IR}$$

In words, the probability gains from exploration over exploitation or no action must exceed the cost differential, normalized by the prize. These conditions are less likely to be met as the cost of exploration rises, but the consideration of cost in players' decision-making is mitigated in tournaments with large prizes that dwarf experimentation costs. As written, they will generate open intervals for $\mu \in \mathbb{R}^+$ in which players will degenerately prefer one of exploration, exploitation, or abandonment. If costs were stochastic – taking a distribution, as is likely the case in practice – the conditions would similarly generate intervals in which one action is more likely than (rather than strictly preferred to) the others.

**Exploration versus Abandonment (IR)**

At what values of $\mu$ are the payoffs to exploration greatest relative to abandonment? I answer this question with the following lemma that characterizes the shape of these payoffs and a proposition establishing the existence of a unique $\mu_1^*$ that maximizes this function.

**Lemma 1.** Payoffs to exploration over abandonment. *The gains to exploration over abandonment are increasing and concave in $\mu$ when $\mu$ is small and decreasing and convex when $\mu$ is large. The gains are zero when $\mu=0$ and approach zero from above as $\mu \longrightarrow \infty$, holding $\beta_1$ fixed.*

**Proposition 1.** *For all values of $q$, there exists a unique level of competition $\mu_1^*$ at which the gains to exploration, relative to abandonment, are maximized.*

According to Lemma 1, a player becomes likely to abandon the tournament when there is either very little competition ($\mu \ll \beta_1$) or very much competition ($\mu \gg \beta_1$). This result constitutes the first empirically testable prediction of the model. The level of competition $\mu_1^*$ at which these gains are greatest is implicitly defined by the following first-order condition:

$$q\left(\frac{-\left(1+\alpha\right)\beta_1}{\left(\left(1+\alpha\right)\beta_1+\mu_1^*\right)^2}\right)+(1-q)\left(\frac{-\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu_1^*\right)^2}\right)+\frac{\beta_1}{\left(\beta_1+\mu_1^*\right)^2}=0$$

**Exploration versus Exploitation (IC)**

I now ask the counterpart question: at what values of $\mu$ are the payoffs to exploration greatest relative to exploitation? I answer this question with a similar lemma and proposition.

**Lemma 2.** Payoffs to exploration over exploitation. *When $q \in (\frac{1}{1+\alpha}, \frac{1}{2})$, the gains to exploration over exploitation are decreasing and convex in $\mu$ for small $\mu$, increasing and concave for intermediate $\mu$, and decreasing and convex for large $\mu$. When $q \in \left(\frac{1}{2}, \frac{3\alpha+1}{4\alpha+1}\right)$, they are increasing and convex for small $\mu$ and decreasing and convex for large $\mu$. When $q > \frac{3\alpha+1}{4\alpha+1}$, they are increasing and concave for small $\mu$ and decreasing and convex for large $\mu$. When $q < \frac{1}{1+\alpha}$, they are decreasing and convex for small $\mu$ and increasing and concave for large $\mu$. In every case, the gains are zero when $\mu = 0$; when $q > \frac{1}{1+\alpha}$ ($q < \frac{1}{1+\alpha}$), they approach zero from above (below) as $\mu \longrightarrow \infty$, holding $\beta_1$ fixed.*

**Proposition 2.** *When $q > \frac{1}{1+\alpha}$, there exists a unique level of competition $\mu_2^*$ at which the gains to exploration, relative to exploitation, are maximized.*

**Corollary.** *When $q < \frac{1}{1+\alpha}$, exploration will never be preferred to exploitation.*

A player's incentive to explore over exploit depends on her relative position in the contest. Provided $q > \frac{1}{1+\alpha}$, in regions where incentive compatibility binds, a player will prefer exploration when she lags sufficiently far behind her competition, and she will prefer exploitation when she is sufficiently far ahead. These results naturally lead to a second empirical prediction: more positive feedback is expected to increase continuing players' tendency to exploit their existing work rather than experiment, but this effect will be offset by greater competition. The level of competition $\mu_2^*$ at which the benefits to exploration are maximized relative to exploitation is defined by the first-order condition for the IC constraint:

$$q\left(\frac{-\left(1+\alpha\right)\beta_1}{\left(\left(1+\alpha\right)\beta_1+\mu_2^*\right)^2}\right)+\left(1-q\right)\left(\frac{-\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu_2^*\right)^2}\right)+\frac{2\beta_1}{\left(2\beta_1+\mu_2^*\right)^2}=0$$

**Tying it together: Exploration vs. the next-best alternative**

**Proposition 3.** *At very low and very high $\mu$, the next-best alternative to exploration is abandonment. At intermediate $\mu$, the next-best alternative is exploitation.*

As $\mu$ increases from zero to infinity, the player's preferred action will evolve from abandonment, to exploitation, to exploration (provided that in expectation it outperforms exploitation, i.e. $q > \frac{1}{1+\alpha}$), to abandonment again. Figure I.1 plots the absolute payoffs to each as the level of competition increases for an example parametrization and highlights each of these regions, holding $\beta_1$ fixed. Note that the region in which players will abandon R&D due to a lack of competition is very narrow, and effectively occurs only with pure monopoly.

[Figure I.1 about here]

Putting the first three propositions together, the implication is an inverted-U shaped effect of competition on innovation: there exists an optimal, intermediate level of competition for motivating experimentation, and it will be attainable as long as experimentation costs are not so large as to make it entirely infeasible for the player. This inverted-U pattern is plotted in Figure I.2 for the same parametrization in Figure I.1.

**Proposition 4.** *When $q > \frac{1}{1+\alpha}$, there exists a unique level of competition $\mu^* \in [\mu_1^*, \mu_2^*]$ at which the gains to exploration are maximized relative to the player's next-best alternative.*

[Figure I.2 about here]

The origins of this result can be traced directly to the IC and IR constraints. Though increasing competition makes experimentation more attractive relative to incremental tweaks, doing so also reduces the gains to continued effort of any kind. At low levels of competition, incentive compatibility binds, such that greater competition increases creative effort. As competition intensifies, the participation constraint eventually binds, and further increases reduce creative effort. Incentives for creativity will generally peak at the point where the participation constraint becomes binding.

At the heart of this model is the player's choice between a gamble and a safe outcome. The concavity of the success function implies that players may prefer the certain outcome to the gamble – forgoing a positive expected quality improvement – even though they are risk-neutral. The inverted-U result is thus robust to risk-aversion, which increases the concavity of payoffs, as well as to limited risk-seeking behavior, provided the utility function does not offset the concavity of the success function.

While these results speak most directly to the incentives of the player with the last move, they carry forward to players with earlier moves. On the one hand, $\mu$ can be equally interpreted as present or anticipated, future competition. The inverted-U pattern will persist even when players internalize competitors' best responses: a player with an inordinate lead or deficit has no reason to continue, one with a solid lead can compel her competitors to abandon by exploiting, and one in a neck-and-neck state or somewhat behind will be most inclined to chance it with exploration to have a fighting chance at winning.

## Remarks and Relation to Previous Literature

The inverted-U effect of competition on experimentation is intuitive. With minimal-to-no competition, the player is already assured victory and will not benefit from additional effort; with extreme competition, the gains to effort are too low to justify continued participation. These patterns are consistent with existing theoretical and empirical results from the tournament literature, which has argued that asymmetries reduce effort from both leaders and laggards (Baik 1994, Brown 2011). The contribution of this model is to consider participation jointly with the explore-exploit dilemma, which adds a new layer to the problem. At intermediate levels of competition, continued participation is justified, but experimentation may not be: with only limited competition, the player is sufficiently well-served by exploiting her previous work. Only at somewhat greater levels of competition will the player have an incentive to experiment.

It is tempting to also draw comparisons against models of patent races, in which firms compete to be the first to arrive at a successful innovation, with the threshold for quality fixed and time of arrival unknown. In innovation contests such as the one modeled here, firms compete to create the highest-quality innovation prior to a deadline. Although Baye and Hoppe (2003) establish an isomorphism between the two, it requires that players are making i.i.d. draws with each experiment. A player's probability of winning in either model is then determined by the number of draws they make – their "effort." This assumption quite clearly does not carry over to the present setting, where designs are drawn from distributions varying across players and over time. Some of the intuition from patent race models nevertheless applies, such as predictions that firms that are hopelessly behind will abandon the competition (Fudenberg et al. 1983).

The model adds a new explanation of an inverted-U pattern to the literature on competition and innovation, and in particular one distinct from that of Aghion et al. (2005), who study the effects of product market competition (PMC) on step-by-step innovation. In the Aghion et al. model, industries can be technologically level or unlevel. In leveled industries, profits are determined by the (exogenous) intensity of price competition in the product market; in unleveled industries, a technological leader earns monopoly rents. When PMC is low, firms tend towards a leveled state, where pre-innovation rents are already large under collusion. When PMC is high, one firm will live in a state of permanent technological leadership, because post-innovation rents are insufficient to motivate the laggard to innovate up to competing in the product market. Incentives for ongoing, back-and-forth innovation by firms are therefore maximized in the middle.

Though the Aghion et al. (2005) result is *prima facie* similar to the one in this section, it is in fact quite different in its theoretical foundations. The primary point of departure is that I study R&D competition for a fixed prize rather than price competition in the product market. In contrast to Aghion et al., competition arises endogenously out of players' choice of whether and how to innovate, and the most competitive contests will be those in which players are technologically similar. Another distinction is the possibility of preemption and leapfrogging in my setting. The two models are thus complementary, in that they show that innovation responds non-monotonically to competition of various types.

## I.2 Graphic Design Contests

I collected a randomly-drawn sample of 122 commercial logo design contests from a widely-used online platform to study competition and the creative process.[15] 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 mock-ups, and many 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 a community of freelance designers (players) in exchange for a fixed prize awarded to its favorite entry. The sponsor publishes a design brief describing 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

[15]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 I.C describes the data collection in detail.

ratings on these designs, but not the ratings on specific competing designs. Copyright is enforced.[16] 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 design brief, which includes a project title and description, 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.[17] 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.[18]

The player identifiers allow me to track each player's 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. Specifically, for every design, I calculate the number of preceding designs in the contest of each rating. I do so both in terms of the prior feedback available (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. I invoke image comparison algorithms commonly used in content-based image retrieval software (similar to Google Image's Search by Image feature) to quantify the originality of each design entered into a contest relative 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 I.C explains exactly how they work. Each one takes a pair of digital images as inputs, summarizes them in terms of a specific, structural feature, and returns a similarity index in [0,1], with a value of one

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

[17] The sponsor may optionally retain the option of not awarding the prize to any entries if none are to its liking.

[18] One of the threats to identification throughout the empirical section is that the effect 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 that I do not observe. This concern is substantially mitigated by evidence from the dataset in Gross (2014), collected from the same platform, in which written feedback is occasionally made publicly available after a contest ends. In cases where it is observed, written feedback is only given to a small fraction of designs in a contest (on average, 12 percent), far less than are rated, and typically echoes the rating given, with statements such as "I really like this one" or "This is on the right track." This written feedback is also not disproportionately given to higher- or lower-rated designs: the frequency of each rating among designs receiving comments is approximately the same as in the data at large. Thus, although the written commentary does sometimes provide players with explicit suggestions or include expressions of (dis)taste for a particular element such as a color or font, the infrequency and irregularity with which it is provided suggests that it does not supersede the role of the 1- to 5-star ratings in practice or confound the estimation in this paper.

indicating a perfect match and a zero indicating total dissimilarity. This index effectively measures the absolute correlation of two images' structural content.

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. This index is my principal measure of experimentation, and it is an empirical counterpart to the parameter $1/\alpha$ in the model. I also make use of related measures: for some specifications, I compare each design against only the best previously-rated designs by the same player or against the best previously-rated designs by competing players. Since players tend to re-use only their highest-rated work, the maximal similarity of a given design to any of that player's previous designs and maximum similarity to her highest-rated previous designs are highly correlated in practice (0.88 for the preferred algorithm, 0.87 for the alternative algorithm).

Creativity can manifest in other ways. For example, players sometimes 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 creative experimentation. 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. Intra-batch similarity is thus an alternative measure that reflects players' tendency to try minor variants of the same concept over a short period of time.

These measures are not without drawbacks or immune to debate. One drawback is that these algorithms require substantial dimensionality reduction and thus provide only a coarse measure of experimentation. 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. One might also question how well these algorithms emulate human perception. The specific examples examined in Gross (2015a) assuage this concern; more generally, I have found these algorithms to be especially good at detecting designs that are plainly tweaks to earlier work (by my perception) versus those that are not, which is the margin that matters most for this paper. Appendix I.C discusses these issues in more detail.

### 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 I.1). On average, 64 percent of designs are rated; less than three receive the top rating. Among rated designs, and the median and modal rating is three stars (Table I.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.[19] The website also provides formal guidance on the meaning of each star rating, which generates consistency in their interpretation and application across different sponsors and contests.

[Table I.1 about here]

[Table I.2 about here]

[19] Another 33 percent of winning designs are rated 4 stars. Twenty-four percent of winning designs are unrated.

Table I.3 characterizes the similarity measures used in the empirical analysis. For each design in the sample, I measure its maximal similarity to previous designs by the same player, previously-rated designs by the same player, and previously-rated designs by the player's competitors (all 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 constituent design files.

[Table I.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 ($\approx$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 these measures.

**Correlations of contest characteristics with outcomes**

To shed light on how these contests operate and how assorted levers affect outcomes of interest, Table I.4 explores the relationship of contest outcomes with prize value, feedback, and other contest characteristics. I borrow the large-sample data of Gross (2015b), which uses a similar (but much larger) sample of logo design contests from the same setting to study the effects of feedback on outcomes of creative tournaments. Though the Gross (2015b) dataset lacks the image files, it includes most of the other variables for these contests. As Appendix I.H shows, this sample is broadly similar to that of the present paper.

The specifications in columns (1) to (3) regress the number of players, designs, and designs per player (as measures of participation) on: the prize value, committed prize value, contest duration, length of the design brief, number of materials provided to be included in the design, and fraction of designs rated. In a departure from existing empirical research on tournaments, these regressions also control for the average cost of effort for all players in the contest as estimated by Gross (2015b), which reflects the design difficulty and would otherwise be an omitted variable biasing the estimated effects of other variables.[20] Column (4) provides estimates from a probit model of whether sponsors of contests with uncommitted prizes choose to award the prize, implying the tournament produced a design good enough to be awarded.

[20]Gross (2015b) develops a semi-parametric procedure to estimate the heterogeneous cost of design for every player in every contest, under the assumption that this cost is constant for a given player in a given contest (as in the model) and the same under both exploration and exploitation (a possibility supported by the model, and a sensible approximation if the variation in cost across players is much larger than the variation in cost of each action for a given player). The procedure effectively uses players' abandonment decision to bound their contest-specific design cost, which must be less than the expected gains from their final design but greater than the gains from an additional design; these gains are estimated in course. Although the average cost in a contest is an imperfect control in that it is calculated from a selected sample of players, it nevertheless appears to be a reasonable estimate of design difficulty, for the reasons discussed in the paper.

[Table I.4 about here]

The estimates in Table I.4 suggest that an extra \$100 in prize value on average attracts an additional 14.8 players, 55.4 designs, and 0.1 designs per player and increases the odds the prize will be awarded by 3.5 percent at the mean of all covariates. There is only a modest and statistically insignificant incremental effect of committed prize dollars, likely because the vast majority of uncommitted prizes in the sample are awarded anyway. Higher-cost contests have lower participation and a lower probability of being awarded. The effects of feedback are equally 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 nearly equal to that of a \$300 increase in the prize – a more than doubling of the average and median prize in the sample.

**Do ratings predict contest success? Estimating the success function**

With the right data, the success function can be directly estimated. Recall from equation (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. This latent value can thus be flexibly specified with fixed effects for each rating (or no rating). The success function can then be structurally estimated as a conditional logit model, using the observed win-lose outcomes of every design in a large sample of contests. To formalize the empirical success function, 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, $\nu_{ijk}$, can be written as follows:

$$\nu_{ijk} = \gamma_{\emptyset}\mathbb{1}(R_{ijk} = \emptyset) + \gamma_1\mathbb{1}(R_{ijk} = 1) + \ldots + \gamma_5\mathbb{1}(R_{ijk} = 5) + \varepsilon_{ijk} \equiv \psi_{ijk} + \varepsilon_{ijk} \qquad \text{(I-6)}$$

This specification is closely related to the theoretical success function in equation (1), with the main difference being a restricted, discrete domain for the feedback. As in the theoretical model, the sponsor is assumed to select as winner the design with the highest value. In estimating the $\gamma$ 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.[21] 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 number of designs by player $j$ in contest $k$, and $I_k$ be the total number of designs entered into that same contest $k$, the empirical success function for player $jk$ takes the following form:

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

Gross (2015b) estimates this model by maximum likelihood using a sample of 496,401 designs entered in 4,294 contests from the same setting. The results are reproduced in Table I.5.

[Table I.5 about here]

[21] I have also tested this assumption by removing subsets of designs from each contest and re-estimating the model. The parameter estimates are statistically and quantitatively similar even when the choice set changes.

Several patterns emerge from this table. The fixed effects are precisely estimated, and the estimated value of a design is monotonically increasing in its rating. Only a 5-star design is on average preferred to the outside option. The contribution of a 5-star design to the success function ($e^{1.53}$) is roughly 12 times that of a 4-star design ($e^{-0.96}$), 137 times that of a 3-star design ($e^{-3.39}$), and nearly 2,000 times that of a 1-star design ($e^{-6.02}$); competition at the top effectively only comes from other 5-star designs. As a measure of fit, the model correctly "predicts" the true winner relatively well, with the odds-on favorite winning almost half of all contests in the sample. These results demonstrate that this simple model fits the data quite well and in an intuitive way, suggesting that ratings provide considerable information about a player's odds of winning. The strong fit of the model also reinforces the assumption that players can accurately assess these odds: though players do not observe the ratings on specific competing designs, they are provided with the distribution of ratings on their competitors' designs, which makes it possible for players to invoke a simple heuristic model such as the one estimated here in their decision-making.

### Empirical Methods and Identification

I exploit variation in the level and timing of the sponsor's ratings to estimate the effects of competition on players' creative choices. With timestamps on all activity, I can determine exactly what a player knows at each point in time about the sponsor's tastes for her work and the competition she faces, and identify the effects of ratings observed at the time of design. Identification is achieved by the quasi-random release of information: it is difficult to predict ex-ante exactly when or how often sponsors will log onto the site to rate new entries, and even more so to predict whether or when a given design will be rated.

Formally, the identifying assumption is that there are no omitted factors correlated with observed feedback that also affect creative choices. This assumption is supported by two pieces of evidence: that (i) the arrival of ratings is unpredictable, and (ii) players' choices are uncorrelated with ratings they have not yet received or cannot observe. The relevant thought experiment is to compare the actions of a player with a 5-star design under her belt before she learns the rating, versus after, or with latent 5-star competition before finding out, versus after – though empirically, undisclosed information is as good as no information.[22]

To establish evidence that feedback provision is unpredictable, I explore the relationship between feedback lags and the rating given. 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 infer the eventual ratings on their designs from the time elapsed without any feedback. Players may also react to uncertainty generated by delays in the provision of feedback, and if this uncertainty is related to the rating given, it would confound my estimates. Table I.6 demonstrates that this is not the case. Column (1) regresses the lag in hours between the time a design is entered and the time it is rated on indicators for the rating given, restricting the sample to designs rated before the contest ends. Column (2) repeats the exercise, measuring the lag as a percent of the total contest duration. Column (3) expands the sample to all rated designs and replaces the dependent variable with an indicator for whether the design was rated prior to the contest's conclusion. I also control for the fraction of the contest elapsed at the time the design was entered, the number

[22] 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 rely on exogenous variation inherent to the setting.

of previous designs by that player and her competitors, and contest and player fixed effects, and cluster standard errors by contest. Across all specifications, I find that lags in feedback provision are unrelated to ratings.

[Table I.6 about here]

Evidence that choices are uncorrelated with unobserved feedback is provided in Section I.3. As a first check, I estimate the effects of observed ratings on experimentation both with and without controls for forthcoming ratings and find the results unchanged. For further evidence, I estimate the relationship between forthcoming ratings and experimentation, finding that with appropriate controls, it is indistinguishable from zero. I similarly examine players' tendency to imitate highly-rated competing designs and find no such patterns – either due to the copyright protection mechanism or, more likely, because players simply do not know which competing designs are highly rated (and therefore which ones to imitate).

## I.3 Competition and the Creative Process

The theoretical predictions can now be put to the test. Section I.3 provides a battery of evidence that conditional on continued participation, competition induces the best-performing players to experiment more than they otherwise would. The basic estimating equation in this part of the paper is the following specification, with variants estimated throughout:

$$\begin{aligned} Similarity_{ijk} \;=\;& \beta_0 + \beta_5 \cdot \mathbb{1}(\bar{R}_{ijk} = 5) + \beta_{5c} \cdot \mathbb{1}(\bar{R}_{ijk} = 5)\mathbb{1}(\bar{R}_{-ijk} = 5) + \beta_{5p} \cdot \mathbb{1}(\bar{R}_{ijk} = 5)P_k \\ &+ \sum_{r=2}^{4} \beta_r \cdot \mathbb{1}(\bar{R}_{ijk} = r) + \gamma \cdot \mathbb{1}(\bar{R}_{-ijk} = 5) + \lambda DR_{ijk} + X_{ijk}\theta + \zeta_k + \varphi_j + \varepsilon_{ijk} \end{aligned}$$

where $Similarity_{ijk}$ is the maximal similarity of design $ijk$ to previous designs by player $j$ in contest $k$; $\bar{R}_{ijk}$ is the highest rating player $j$ received in contest $k$ prior to design $ijk$; $\bar{R}_{-ijk}$ is the highest rating player $j$'s competitors received prior to design $ijk$; $P_k$ is the prize in contest $k$ (measured in \$100s); $DR_{ijk}$ is the number of days remaining in the contest at the time design $ijk$ is entered; $X_{ijk}$ is a vector of design-level controls; and $\zeta_k$ and $\varphi_j$ are fixed effects.

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. In the second set, I replace the dependent variable with the similarity to that player's best, previously-rated designs, and then within-batch similarity. The third set of regressions examines the change in similarity to previously-rated designs, as a function of newly-received feedback. The fourth set of regressions tests the aforementioned identifying assumption that players are not acting on private information that I cannot observe. The fifth set of regressions tests whether players imitate high-performing competitors, which they should not be able to discern from the information they are given.

Section I.3 provides the counterpart analysis examining the effects of competition on players' tendency to continue investing in the contest. The evidence substantiates the model's second prediction: that increasing competition can drive players to quit. The specifications in this section are similar to those of the experimentation regressions. I estimate variants of the following model:

$$\begin{aligned} Abandon_{ijk} &= \beta_0 + \sum_{r=1}^{5} \beta_r \cdot \mathbb{1}(\bar{R}_{ijk} = r) + \sum_{r=1}^{5} \gamma_r \cdot \mathbb{1}(\bar{R}_{-ijk} = r) \\ &\quad + \sum_{r=1}^{5} \delta_r \cdot \mathbb{1}(\bar{R}_{ijk} = r) N_{-ijk} \\ &\quad + \delta N_{-ijk} + \lambda DR_{ijk} + X_{ijk}\theta + \zeta_k + \varphi_j + \varepsilon_{ijk} \end{aligned}$$

where $Abandon_{ijk}$ indicates that player $j$ entered no additional designs in contest $k$ after design $ijk$; $N_{-ijk}$ is the number of five-star designs by player $j$'s competitors in contest $k$ at the time of design $ijk$; and $\bar{R}_{ijk}$, $\bar{R}_{-ijk}$, $DR_{ijk}$, $X_{ijk}$, $\zeta_k$, and $\varphi_j$ retain their previous definitions. The precise moment at which each player makes an active choice to abandon is impossible to measure, and I thus use inactivity as a proxy. In general, this measure does not distinguish between a "wait and see" approach that ends with abandonment versus abandonment immediately following design $ijk$. Since the end result is the same, the distinction is immaterial for the purposes of this paper. Note that standard errors throughout both Sections I.3 and I.3 are clustered by player to account for any within-player correlation in the error term.

## Competition and Experimentation

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

I begin by studying players' tendency to tweak any of their previous work in a contest. Table I.7 provides estimates from regressions of the maximal similarity of each design to previous designs by the same player on indicators for the highest rating that player had previously received. All specifications include interactions of the indicator for having received the top rating with (i) the prize value (in \$100s) and (ii) a variable indicating the presence of top-rated competition, as well as contest and player fixed effects. The even-numbered columns additionally control for the fraction of the contest elapsed at the time of submission and the number of designs previously entered by the player and her competitors, which characterize the overall state and progression of the contest. Columns (3) and (4) control for future feedback on the player's earlier work; if players have contest-specific ability or other information unobserved by the researcher (e.g., sponsors' written comments), it will be accounted for by these regressions.

[Table I.7 about here]

The results are consistent across all specifications in the table. Players with the top rating enter designs that are 0.3 points, or roughly one full standard deviation, more similar to previous work than players who have low (or no) feedback. Roughly one third of this effect is reversed by high-rated competition. With a highest observed rating of four stars, new designs are on average around 0.1 points more similar to previous work. This effect further attenuates as the best observed rating declines, and it is indistinguishable from zero at a best observed rating of two stars.

In practice, players tend to tweak only their best work. Table I.8, columns (1) and (2) estimate a variant on the first two columns of Table I.7, regressing each design's maximal similarity to the *highest-rated* preceding designs by the same player on the same explanatory variables. Columns (3) and (4) use a sample of batches and the alternative measure of experimentation: the maximal similarity of any two designs in each batch. Columns (5) and (6) repeat the latter exercise, weighting observations by batch size. All specifications control for contest and player fixed effects, and the table shows variants of the regressions with and without design- and batch-level covariates.

[Table I.8 about here]

The results for the design-level regressions (Columns 1 and 2) are similar to but slightly stronger than those of the previous table. Players with the top rating enter designs that are 0.35 points, or about 1.3 standard deviations, more similar to their highest-rated work in that contest, but this effect is reduced by more than half when there is top-rated competition, and the tendency tweak is again monotonically decreasing in the highest rating the player has received.

Columns (3) to (6) confirm that competition has similar effects on experimentation within batches. When entering multiple designs at once, the maximal similarity of any two designs in the batch declines 0.3 points, or approximately one standard deviation, for players with a top rating who also face top-rated competition, relative to those who do not. Top players facing competition are thus more likely to experiment not only across batches but also within them. The consistency of the results demonstrates that they are not sensitive to inclusion of controls or weights.

The regressions in Tables I.7 and I.8 use contest and player fixed effects to control for factors that are constant within contests, across players or within players, across contests, but they do not control for factors that are constant throughout a given contest for a given player, as doing so leaves too little variation for me to identify the effects of feedback and competition. Such factors may nevertheless be confounding omitted variables. For example, if players can sense their match to a particular contest, and change their behavior accordingly throughout the contest, the estimated effects may be confounded by this unobserved self-selection – though such concerns are in part relieved by the consistency of results in Table I.7 controlling for forthcoming ratings. The estimates in the previous tables additionally mask potential heterogeneity that may be present in players' reactions to feedback and competition over the course of a contest.

Table I.9 addresses these issues with a model in first differences. The dependent variable is the change in 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 experimenting to copying; and a value of -1, the converse. The independent variables are changes in indicators for the highest rating the player has received, with the usual interactions of the top rating with the prize and the presence of top-rated competition. I estimate this model with assorted configurations of fixed effects and controls to account for other reasons why experimentation may vary over time, though the results are not statistically different across specifications.

[Table I.9 about here]

The results provide the most powerful evidence thus far on the effects of feedback and competition on experimentation. When a player receives her first five-star rating, her next design will be a near replica. The similarity score increases by nearly 0.9 points, or *three* standard deviations. Top-rated competition shaves nearly half of this effect. Given their magnitudes, these effects will be plainly visible to the naked eye. The effects of new best ratings of four-, three-, and two-stars on experimentation attenuate monotonically, akin to previous results.

Interestingly, these regressions also find that new recipients of the top rating can be induced to experiment by larger prizes. The theory suggests a natural explanation: large prizes moderate the role of costs in players' decision-making. If experimentation is more costly (takes more time or effort) than tweaking, it may only be worth doing when the prize is large. This is particularly the case for players with highly-rated work in the contest, given how the shape of and movement along a player's success function depends on the quality of her prior submissions.

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

This latter result is especially important for ruling out an information-based story. The fact that other designs received a 5-star rating might signal that the sponsor has diverse preferences and that experimentation has a higher likelihood of success than the player might otherwise believe. If this were the case, we should see experimentation continue to rise as 5-star competition grows. That this is not the case suggests that the effect is in fact the due to incentives.

In unreported regressions, I look for effects of five-star competition on experimentation by players with only four-star designs, and find attenuated effects that are negative but not significantly different from zero. I also explore the effect of prize commitment on experimentation, since the sponsor's outside option of not awarding the prize is itself a competing alternative – one which according to the conditional logit estimates in Table I.5 is on average preferred to all but the highest-rated designs. The effect of prize commitment is not estimated to be different from zero. I similarly test for effects of the presence of four-star competition on experimentation by players with five-star designs, finding none. These results reinforce the perception that competition in effect comes from designs with the top, five-star rating.

**Similarity of new designs 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 I.10 test whether players' creative choices are related to forthcoming, not-yet-available feedback. If an omitted determinant of creative choices is correlated with the feedback, then it would appear as if experimentation responds to future ratings, but if the identifying assumptions hold, I should only find zeros.

[Table I.10 about here]

The specification in Column (1) regresses a design's maximal similarity to the player's best designs that will eventually be – but have not yet been – rated on indicators for the ratings they later receive.

The estimates ostensibly suggest a potential failure of the identifying assumptions: although many are not significantly different from zero, the point estimates imply that players tweak these "placebo best designs" that have yet to be rated more or less depending on the rating they eventually receive, and that competition continues to induce experimentation, suggesting that it's not feedback per se that shapes creative choices, but rather some omitted factors that correlate with it. However, similarity to a high-rated placebo may in fact be the result of tweaks on an even earlier design that the placebo also happens to look like. Column (2) of the table thus controls for both the given and placebo designs' similarity to the *observed* best design at the time; Column (3) relaxes these controls to vary by the observed best rating. As a final check, I isolate the similarity to the placebo best design that cannot be explained by similarity to a 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 positive forthcoming ratings. Feedback only relates to creative choices when it is observed at the time of design.

**Imitation of competing designs**

Though players can see competing designs in the same contest, they see only the distribution of feedback these designs have received – not the ratings on specific, competing entries – and should therefore not be able to use this information to imitate highly-rated competitors. The regressions in Table I.11 test this assumption by examining players' tendency to imitate competitors.

The first two columns of the table provide estimates from regressions of similarity to the highest-rated design by competing players on indicators for its rating. As in previous specifications, the top-rating indicator is interacted with the prize and with an indicator for whether the player herself also has a top-rated design in the contest. The latter columns repeat the exercise with first-differenced variants of the same specifications. There is little evidence in this table that players imitate highly-rated competitors in any systematic way – likely because they are simply unable to identify which competitors are highly-rated. In unreported results, I replace the left-hand side with imitation of *any* competing design and similarly find no effect. The results establish that "experimentation" in the presence of competition is not just imitation of competitors' designs. Appendix Table I.D.5 provides counterpart estimates using the difference hash algorithm, which suggest that if anything, players tend to deviate *away* from competitors' high-rated work.

[Table I.11 about here]

## Competition and Abandonment

Having established that competition induces players to experiment, it remains to be seen how competition affects players' decision to continue in versus abandon a contest. In Table I.12 I examine the effect of a player's first rating and the competition she faces when it is received on the probability she subsequently enters at least one more design. I focus on the first rating a player receives because it will typically be ex-ante unpredictable. The specifications in the table regress this measure of abandonment on dummies for each rating the player may have received, alone and interacted with the number of top-rated competing designs, the latter as a distinct regressor, and dummies for the highest rating on competing designs at the time.

[Table I.12 about here]

Columns (1) to (3) estimate linear specifications with contest, player, and contest and player fixed effects. Linear specifications are used in order to control for these fixed effects (especially player fixed effects), which may not be estimated consistently in practice and could thus render the remaining estimates inconsistent in a binary outcome model, due to an incidental parameters problem. Column (4) estimates a logit model with only the contest fixed effects. The linear model with two-way fixed effects (in Column 3) is the preferred specification.

Players with poor initial feedback drop out with probability close to one. Those with high initial feedback enter additional designs with roughly a 50 percentage-point higher rate, but competition counteracts this effect: with just a handful top-rated competitors, a player is likely to walk away no matter what rating she receives. The effect is significant at only the 10 percent level, and thus somewhat imprecise. But it appears that by driving players to stop investing, heavy competition can discourage experimentation just as much as an absence thereof.[23]

I also study abandonment at points in a contest other than immediately following a player's first rating. Table I.13 estimates the probability that a given design is a player's final design on the feedback and competition observed at the time. As previously discussed, this measure could reflect either a simulatneous choice to stop investing or a "wait and see" strategy that yields no further action – although according to one designer who participates on this platform, it is often the case that players will enter their final design knowing it is such and never look back.

[Table I.13 about here]

This table again estimates three linear specifications and a logit model, with the same arrangement of fixed effects, and adding the design-level controls from earlier sections. The independent variables are analogous to those in the previous table, measured at the time the given design was submitted. In the preferred, linear specification of Column (3), I find that players with a top-rated design are more likely to subsequently enter more designs, but this effect is negated by the presence of one five-star competitor, and more than offset by multiple five-star competitors – with all effects significant at the one percent level.

## I.4 Does it Really Pay to Innovate?

Why do the designers in these contests respond to competition by experimenting with new ideas? In conversations with creative professionals (including the panelists hired for the exercise below), many have asserted that competition means that they need to "be bold" or "bring the 'wow' factor," and that it induces them to take creative risks. Gambling on a more radical, untested idea is thus a calculated and intentional choice. The implicit assumption motivating this type of creative risk-taking both in the model and in practice is that experimentation is a high-risk, high-return endeavor – the upside to experimentation is what makes it worthwhile. This assumption is pervasive not only in research, but also in the public discourse on innovation and entrepreneurship. Whether or not it is true is ultimately an empirical question.

A natural way to answer this question in the present context is to examine the distribution of sponsors' ratings on radical versus incremental designs in the sample. To do so, I categorize

[23] Note that although this outcome runs counter to the interests of the principal, it may be desirable from a social welfare perspective if the high-rated designs are unlikely to be outdone. See Appendix I.G for further discussion.

designs as tweaks if they have similarity to any earlier designs by the same player of 0.7 or higher and record the rating of the design they are most similar to; I classify designs as experimental if their maximal similarity to earlier designs by that player is 0.3 or below and record the highest rating the player had previously received.[24] I then compute the distribution of sponsors' ratings on this subsample, conditioning on the rating of the tweaked design (for tweaks) or the highest rating previously given to that player (for experimentation).

Figure I.3 illustrates these distributions. Although the modal rating in all cases is that of the conditioning variable, the figure demonstrates that experimentation is indeed higher variance than tweaking. Experimenting after poor feedback on average outperforms tweaks to designs with low ratings, especially considering that the top rating is orders of magnitude more valuable to a player than lower ratings (Table I.5). Yet experimentation appears to on average *underperform* tweaks of top-rated designs, raising the question of why a player would deviate from her top-rated work.

[Figure I.3 about here]

The problem with this analysis is that the observed outcomes are censored: it is impossible to observe the fruits of experimentation beyond a five-star rating. With this top-code in place, exploration after a five-star design will necessarily appear to underperform exploitation – in the data, the sponsor's rating can only go down. The data are thus inadequate for evaluating the benefits to experimentation for players at the top. To circumvent the top-code, I hired a panel of professional graphic designers to independently assess all designs in my sample that were rated five stars by contest sponsors, and I look to the panelists' ratings for evidence that experimentation is in fact high-risk, high-return.

### Results from a Panel of Professional Designers

To obtain independent appraisals of all 316 five-star designs in the sample, I hired five professional graphic designers at their regular rates to administer their own ratings to each design on an extended scale. These ratings were collected though a 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 (as excerpted from the source data) and asked to rate the design's "quality and appropriateness" on a scale of 1–10.

Appendix I.F provides more detail on the survey procedure and shows the distribution of ratings from each panelist. One panelist ("Rater 5") was a particularly critical judge and frequently ran up against the lower bound. The mass around the lower bound was apparent after the first day of the survey, and though I include this panelist in the appendix for disclosure, the decision was made at that time to exclude these ratings from subsequent analysis. The results are nevertheless robust to including ratings from this panelist above the lower bound.

To account for differences in the remaining panelists' austerity, I first normalize their ratings by demeaning, in essence removing rater fixed effects. For each design, I then compute summary statistics of the panelists' ratings (mean, median, maximum, and s.d.). As an alternative approach to aggregating panelists' ratings, I also calculate each design's score along the first principal component generated by a principal component analysis. Collectively, these summary statistics

[24] A player's intentions are more ambiguous at intermediate values, which I accordingly omit from the estimation.

characterize the distribution of opinion on a given design. One way to think about them is as follows: if contest sponsors were randomly drawn from this population, then the design's realized rating would be a random draw from this distribution.

I identify designs as tweaks or experimentation using the definitions above and then compare the level and variation in panelists' ratings on designs of each type. Table I.14 provides the results. Designs classified as tweaks are typically rated below-average, while those classified as experimentation 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 of the panelists is also greater for experimentation, with the difference significant at the one percent level. Yet so is the level of disagreement: the standard deviation across panelists' ratings on a given design is significantly greater for experimentation than for tweaks. The evidence thus appears to support the popular contention that radical innovation is both higher mean and higher variance than incremental innovation, even at the top.[25]

[Table I.14 about here]

## I.5 When is Experimentation Most Likely?

The reduced-form results establish that while competition can motivate high performers to experiment with new ideas, too much competition will drive them out of the market altogether. How much is "too much"? Given that the full effect of competition on the degree of experimentation is achieved by a single, high-quality competitor, and that players are increasingly likely to quit as competition intensifies, it would be natural to conclude that incentives for active experimentation peak in the presence of exactly one top-rated competitor – just enough to ensure that competition exists without further eroding the returns to effort.

To formalize an answer to this question, I estimate a choice model in which with each submission, a player selects from the three basic behaviors I observe in the data: (i) tweak and enter more designs, (ii) experiment and enter more designs, and (iii) do either and subsequently abandon the contest. To distinguish between players who are more likely to be truly giving up versus adopting a "wait and see" strategy, I condition the latter case on the player's contemporaneous probability of winning, calculated using the conditional logit estimates in Table I.5. This model will allow me to determine on which margin players are operating as competition builds and to identify the conditions under which active experimentation is most likely. As before, I classify each design as a tweak if its similarity to any earlier design by the same player is 0.7 or higher and an experiment if its maximal similarity to earlier designs by that player is 0.3 or lower.

Each action in this choice set is assumed to have latent utility $u_{ijk}^{a}$, where $i$ indexes submissions by player $j$ in contest $k$. I model this latent utility as a function of the player's own ratings, com-

[25] If anything, these differences may be understated. If a player enters the same design twice, the first would be classified as an experiment, and the second as a tweak, but they would receive the same rating from panelists. Excluding designs that are either tweaks of or tweaked by others in the sample does not affect the results.

petitors' ratings, the time remaining in the contest, additional controls, and a logit error:

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

The explanatory variables are defined as before: $\bar{R}_{ijk}$ is the highest rating player $j$ has received in contest $k$ prior to $ijk$, $\bar{R}_{-ijk}$ is the highest rating on competing designs, $N_{-ijk}$ is the number of top-rated competing designs, $DR_{ijk}$ is the number of days remaining, and $X_{ijk}$ are controls.

I estimate this model and then use the results to predict the probability that a player with a 5-star design 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.[26] These probabilities are shown in Figure I.4. Panel A plots the probability that the player tweaks and enters another design; Panel B, that she experiments and enters another design; and Panels C and D, that she abandons, conditional on her probability of winning at that time being 0.5 versus 0.05.[27] The bars around each point provide the associated 95 percent confidence interval.

[Figure I.4 about here]

The probability that a player tweaks and remains active (Panel A) peaks at 52 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 actively experiments (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). Panels C and D show that the probability of abandonment increases monotonically in the level of competition, and approaches 80 percent for players with a low probability of success.

Observed behavior thus appears to conform to the predictions of economic theory: when competition is low, players are on the margin between exploration and exploitation, whereas when competition is high, they straddle the margin between exploration and abandonment. The results of this execise also agree with the reduced-form evidence, in finding that high-rated players are most likely to actively experiment when they encounter precisely one highly-rated competitor.

[26] For the case of no 5-star competitors, I assume the highest rating on any competing design is 4 stars.

[27] To produce Panels C and D, I estimate a choice model that includes this probability as an explanatory variable. The results are not sensitive to this choice, which is taken in order to separate players who are competitive and those who lag far behind. In unreported estimations, I also split abandonment into "tweak and abandon" and "experiment and abandon." The exercise reveals that players rarely tweak and then abandon, and the probability of doing so is statistically invariant to competition, but the probability that players experiment and then abandon is significantly increasing in competition. This evidence is strikingly consistent with the theoretical model, which suggests that tweaking and abandoning isn't a margin where we should see much activity, and that players who abandon will be on the margin with experimentation.

Panel C directly illustrates the inverted-U shaped effect of competition on experimentation and is an empirical counterpart to the theoretical illustration in Figure I.2.

## I.6 Implications for Research, Management, and Policy

These results have direct implications for policies and programs to incentivize innovation, in both the workplace and the market. The foremost result is that the sharp incentives of prize competition can motivate creative effort in a work environment, but that doing so requires striking a delicate balance in the intensity of competition. In designing contracts for creative workers, managers would be keen to offer incentives for high-quality work relative to that of peers or colleagues, in addition to the traditional strategy of establishing a work environment with intrinsic motivators such as intellectual freedom, flexibility, and challenge. Another advantage of the tournament-style incentive structure is that it incorporates tolerance for failure by allowing players to recover from unsuccessful experimentation, which has been shown to be an important feature of contracts for motivating innovation (e.g., Manso 2011, Ederer and Manso 2013).

In practice, the 'Goldilocks' level of competition preferred by a principal may be difficult to achieve, much less determine. Finding it could potentially require experimentation with the mechanism itself, such as by changing the prize; subsidizing or restricting entry; or eliminating non-preferred players midway through the contest. In this paper, one high-quality competitor was found to be sufficient to induce another high-quality player to experiment, and further increases in competition have the effect of driving players away. As a rule of thumb for other settings, a good approximation may be to assume that one competitor of equal ability is enough to induce innovative effort, but having more than a few such competitors is likely more harmful than helpful.

The results also have bearing on design of public incentives for R&D, which is itself a creative endeavor, and the implementation of other policies (such as antitrust policy) undertaken with the intent of incentivizing innovation. Although this paper is fundamentally about individuals, the theoretical framework can be interpreted as firms competing in a winner-take-all market. This interpretation is not without some peril, as markets are inherently more dynamic and less structured than the model allows.[28] The results nevertheless shed light on the forces that define the relationship between competition and innovation, particularly in settings where post-innovation rents are much larger than competitors' pre-existing rents.

Three concrete policy implications follow. The first is support for prize competition as a mechanism for generating innovation. While the focus of this paper is graphic design for marketing materials, it is conceivable to think that similar forces might be at work in other creative endeavors, including R&D. For example, Scotchmer (2004) recounts that in the 1970s, the U.S. Air Force established a system whereby rival companies vying for fighter jet contracts would build prototypes and fly them in competition to demonstrate quality, with the top performer winning a production contract – a process which ultimately led to the F-16 and F-18 fighter jets. Looking further back in history, Brunt, Lerner, and Nicholas (2012) show that prize competitions sponsored by the Royal Agricultural Society of England in the 19th and 20th centuries generated subsequent innovation

[28] In many settings, incentives for innovation are shaped by both post- and pre-innovation rents, the latter of which are absent in this paper. The model could be extended to account for pre-innovation rents by re-defining the "prize" as the incremental rents and allowing it to vary by player – though even with this modification, many of the features of sequential innovation in other settings would be absent from the model.

in the targeted areas, and Moser and Nicholas (2013) show that prizes offered at the 1851 Crystal Palace Exhibition shaped the direction of inventive activity for years to come. These issues are particularly relevant today, as governments, private foundations, and firms commit ever larger sums to R&D prizes and institutionalize prize competition.[29] The U.S. government now even operates a website where federal agencies can run public prize competitions for problems large and small, and has hosted over 280 contests from nearly 50 agencies with prizes ranging from status only (non-monetary) to tens of millions of dollars (OSTP 2014).

The second implication is an argument for monopoly and perfect competition potentially being equally harmful to innovation in market settings. According to the most recent U.S. Horizontal Merger Guidelines (2010), "competition often spurs firms to innovate," and projected post-merger changes in the level of innovation is one of the government's criteria for evaluating mergers. The results of this paper suggest that a transition from no competition to some competition increases incentives for radical innovation over more modest, incremental improvements to existing technologies, but that the gains to innovation can decline to zero in crowded or overly competitive markets, leaving participants content to remain with the status quo.

A final implication of the results in this paper is that contrary to the conventional wisdom that duplicated R&D is purely wasteful,[30] simultaneous, duplicated efforts may be ex-ante efficient: the competition of a horserace may induce more radical innovation, whose fruits might compensate for the deadweight loss of the duplicated effort. From a social welfare perspective, institutional policies prohibiting joint support of dueling research programs would then do more harm than good. This corollary requires further testing, but if true, it suggests not only a fresh look at existing research on the welfare impacts of R&D, but potentially important changes to both R&D policy and strategies for managing innovation in the firm.

## I.7 Conclusion

Ingenuity undoubtedly occurs along a continuum, with some innovations being inherently more novel than others. Consider the smartphone: the first Apple iPhone was extremely original at the time it was developed, while later generations and competitors have essentially only tweaked the design with hardware and operating system changes. Though most innovation consists of modest, incremental advances, many historically important innovations were more radical departures from the status quo. Understanding what inspires individuals to experiment with new and untested ideas is thus critical to policy and management practices implemented to foster innovation.

This paper combines theory, data, and new tools for measuring experimentation to show that while some competition is necessary to induce high performers to experiment with new ideas, excessive competition can equally discourage innovative effort. The results imply that there is an intermediate level of competition that maximizes incentives for innovation. In the setting of this paper, this intermediate value is exactly one high-performing competitor.

These results tie together the literatures in bandit decision models and tournament competition, and they provide what is to my knowledge the most direct evidence yet available on how incentives affect experimentation within the creative process. The results also contribute to a long-standing debate

[29]See Williams (2012) for more examples and an in-depth review of the literature on innovation inducement prizes.
[30]For example, see Jones (1995), Jones and Williams (1998), and Jones and Williams (2000).

on the effects of competition on innovation that dates back to Schumpeter (1942). Previous research has returned evidence of positive, negative, and inverted-U relationships between competition and innovation but generally suffers from inconsistencies and imprecision in measuring competition and innovation, lack of econometric identification, and confusion regarding the economic mechanism at play. This paper addresses these issues by establishing clear and precise measures of key quantities, exploiting the arrival of information on the state of competition to identify its effects, and clarifying the mechanism responsible for the results. The end result is clear-cut evidence of an inverted-U effect of competition on innovation in winner-take-all markets.

Many questions and opportunities remain for future research. Most importantly, as Weitzman (1996) writes, "we need to understand, much better than we do, the act of human innovation." Is the essence of innovation the recombination of existing ideas in new forms, or the creation of something truly new? What are the implications for the increasingly unpopular classical liberal arts education, which exposes students to diverse views and approaches to problem-solving, versus specialized training? Can diversity in teams compensate for a lack of breadth within its individual members? Another goal for future research is to better understand how the creative process unfolds, and especially how it adapts to constraints. A final question is whether successful innovation is stochastic, deterministic in research inputs, or something in between, as the answer has direct implications for how innovation is modeled or measured in other settings.

Figure I.1: Payoffs to each of exploration, exploitation, and dropout (example)

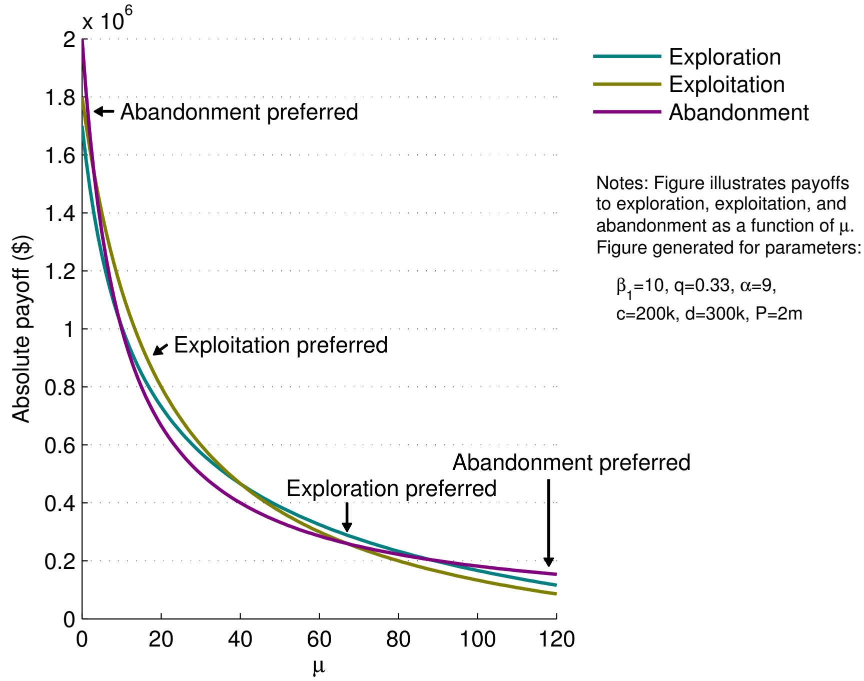



Figure I.2: Payoff to exploration over next-best option (example)

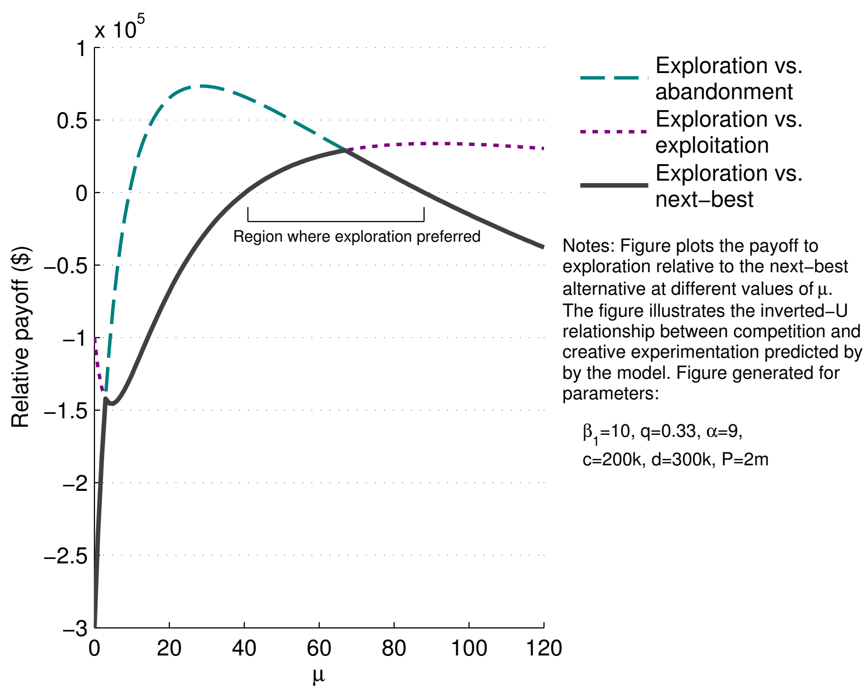
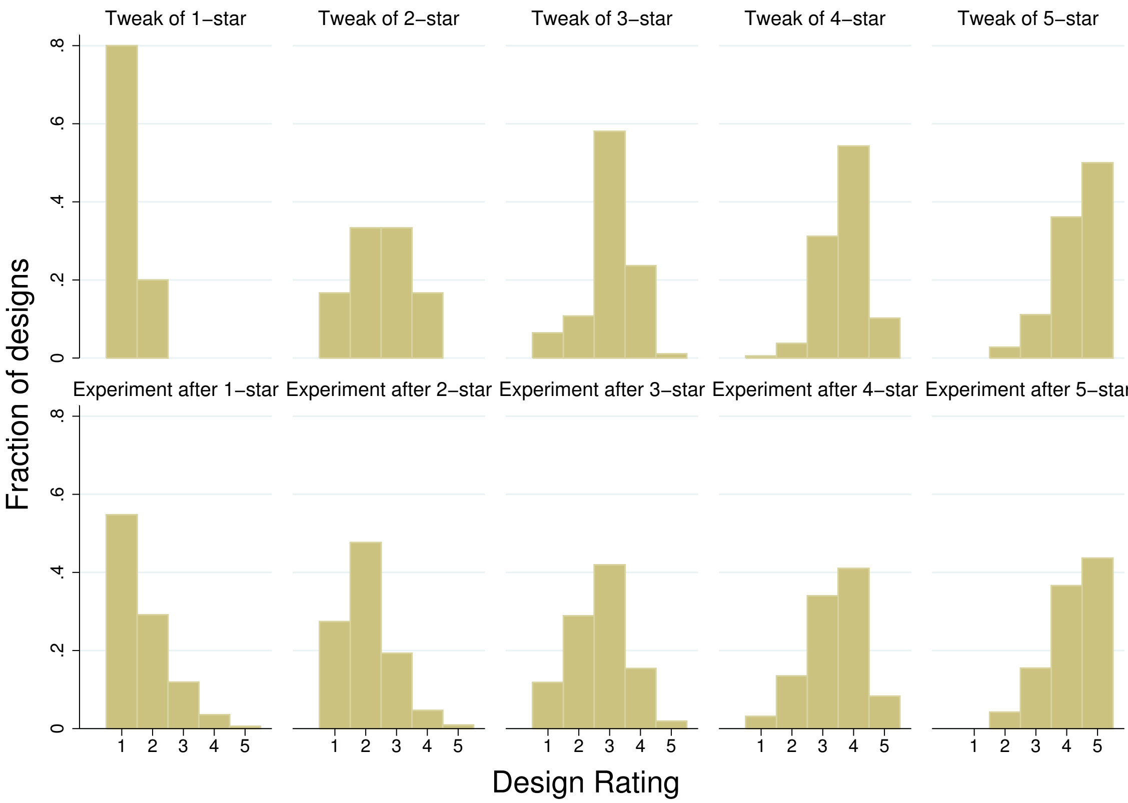
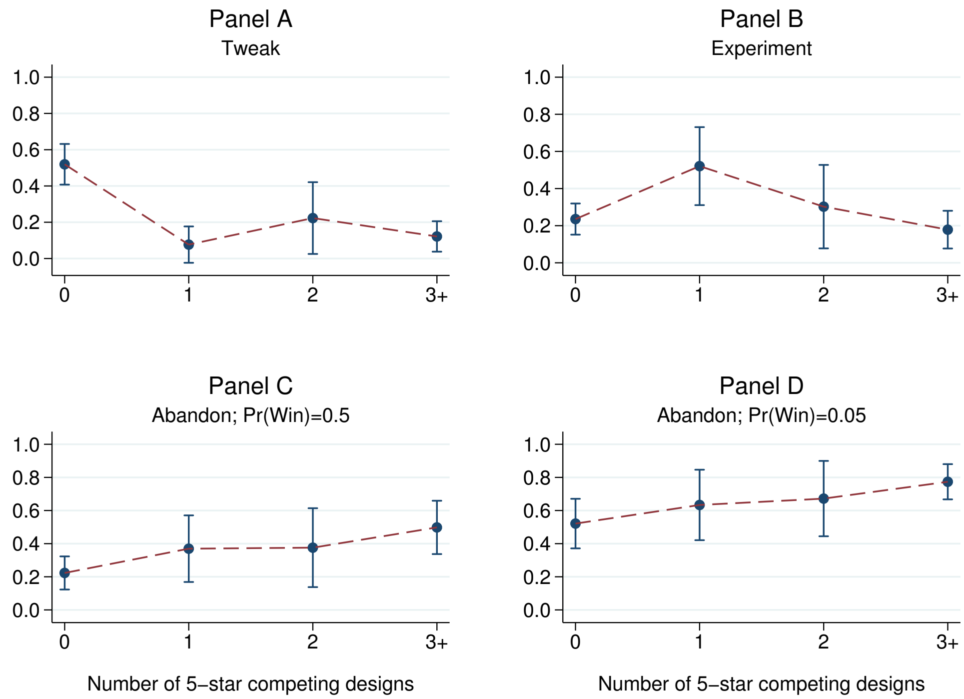


Figure I.3: Sponsor ratings on tweaks vs. experimental designs




Notes: Figure shows the distribution of ratings given to tweaks and experimental designs. Each design in the sample is classified as a tweak if its maximal similarity to any previous design by the same player is greater than 0.7 and experimental if less than 0.3. This figure uses the perceptual hash algorithm to calculate similarity scores. Sample size in each subfigure, from left-to-right across each row: 10, 24, 93 (first row); 186, 36, 1828 (second row).

Figure I.4: Probability of tweaking, experimenting, and abandonment as a function of competition




Notes: The figure plots the probability that a player who already has at least one 5-star rating in a contest does one of the following on (and after) a given submission: tweaks an existing design and then enters more designs (Panel A), experiments and then enters more designs (Panel B), and does either and subsequently abandons the contest, as a function of her contemporaneous probability of winning (Panels C and D). These probabilities are estimated as described in the text, and the bars around each point provide the associated 95 percent confidence interval.

The figure establishes that active experimentation is equally non-monotonic over competition in practice as it is in the theoretical model. Panel B directly illustrates this inverted-U pattern. This non-monotonicity appears to arise for the posited reasons: when competition is low, players are on the margin between tweaks and experimentation (the incentive compatibility constraint, Panels A and B); as competition increases, they are increasingly likely to stop investing, especially when their probability of winning is very low (participation constraint, Panels C and D).

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

| 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 |
| Image missing | 11,758 | 0.04 | 0.19 | 0.00 | 0.00 | 0.00 |
| **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 |
| Image missing | 11,758 | 0.04 | 0.19 | 0.00 | 0.00 | 0.00 |

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). 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 I.4: Correlations of contest outcomes with their characteristics

| | (1) Players | (2) Designs | (3) Designs/Player | (4) Awarded |
|---|---|---|---|---|
| Total Prize Value ($100s) | 14.828*** | 55.366*** | 0.124*** | 0.248*** |
| | (0.665) | (2.527) | (0.015) | (0.042) |
| Committed Value ($100s) | 1.860* | 5.584 | 0.008 | |
| | (1.118) | (4.386) | (0.025) | |
| Average Cost ($) | -1.790*** | -9.074*** | -0.088*** | -0.133*** |
| | (0.096) | (0.353) | (0.004) | (0.010) |
| Fraction Rated | -14.276*** | -20.056*** | 0.683*** | 0.691*** |
| | (0.812) | (2.855) | (0.040) | (0.106) |
| Contest Length | 0.340*** | 1.113*** | 0.003 | 0.007 |
| | (0.069) | (0.251) | (0.004) | (0.010) |
| Words in Desc. (100s) | 0.061 | 2.876*** | 0.059*** | -0.158*** |
| | (0.081) | (0.389) | (0.005) | (0.014) |
| Attached Materials | -0.878*** | -1.557** | 0.051*** | -0.011 |
| | (0.161) | (0.604) | (0.012) | (0.016) |
| Prize Committed | 1.076 | 2.909 | -0.023 | |
| | (3.290) | (12.867) | (0.085) | |
| Constant | 9.150*** | -4.962 | 2.488*** | 1.967*** |
| | (1.760) | (6.180) | (0.073) | (0.179) |
| N | 4294 | 4294 | 4294 | 3298 |
| $R^2$ | 0.63 | 0.65 | 0.31 | |

Notes: Table shows the estimated effect of contest attributes on overall participation and the probability that the prize is awarded from Gross (2015b), controlling for the average cost of participating players. 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.

Table I.5: Conditional logit of win-lose outcomes on ratings

| **Fixed effect** | **Est.** | **S.E.** | **t-stat** | **Corresponding $\beta$** |
|---|---|---|---|---|
| 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 estimates from conditional logit estimation of the win-lose outcome of each design as a function of its rating. Outside option is not awarding the prize, with utility normalized to zero. The design predicted by the model as the odds-on favorite wins roughly 50 percent of contests.

Table I.6: Correlation of feedback lags with the rating given

| | (1) Lag (hours) | (2) Lag (pct. of contest) | (3) Rated before end? |
|---|---|---|---|
| Rating==5 | 1.577 | 0.006 | -0.023 |
| | (3.394) | (0.016) | (0.031) |
| Rating==4 | -2.389 | -0.013* | 0.009 |
| | (1.727) | (0.007) | (0.020) |
| Rating==3 | -0.740 | -0.004 | 0.007 |
| | (2.106) | (0.009) | (0.016) |
| Rating==2 | 1.167 | 0.005 | 0.006 |
| | (1.904) | (0.008) | (0.011) |
| Constant | 16.655* | 0.134*** | 1.075*** |
| | (8.514) | (0.042) | (0.050) |
| N | 7388 | 7388 | 8144 |
| $R^2$ | 0.45 | 0.48 | 0.63 |
| Controls | Yes | Yes | Yes |
| Contest FEs | Yes | Yes | Yes |
| Player FEs | 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 time between submission and feedback, in hours; Column (2), this lag as a fraction of the contest length; and Column (3), an indicator for whether a design receives feedback before the contest ends. All columns control for 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.

Table I.7: Similarity to player's previous designs

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Player's prior best rating==5 | 0.284*** | 0.277*** | 0.289*** | 0.286*** |
| | (0.085) | (0.087) | (0.085) | (0.087) |
| * 1+ competing 5-stars | -0.106* | -0.118** | -0.101* | -0.112* |
| | (0.058) | (0.059) | (0.058) | (0.058) |
| * prize value ($100s) | -0.013 | -0.021 | -0.011 | -0.020 |
| | (0.028) | (0.028) | (0.028) | (0.029) |
| Player's prior best rating==4 | 0.099*** | 0.077*** | 0.111*** | 0.090*** |
| | (0.017) | (0.017) | (0.017) | (0.018) |
| Player's prior best rating==3 | 0.039*** | 0.029** | 0.050*** | 0.039*** |
| | (0.014) | (0.014) | (0.014) | (0.014) |
| Player's prior best rating==2 | -0.004 | -0.009 | 0.007 | 0.001 |
| | (0.020) | (0.020) | (0.020) | (0.020) |
| One or more competing 5-stars | -0.014 | -0.016 | -0.016 | -0.018 |
| | (0.020) | (0.022) | (0.019) | (0.022) |
| Days remaining | -0.005* | -0.009 | -0.005** | -0.009 |
| | (0.003) | (0.007) | (0.003) | (0.007) |
| Constant | 0.351* | 0.414** | 0.355* | 0.419** |
| | (0.181) | (0.195) | (0.181) | (0.196) |
| N | 5075 | 5075 | 5075 | 5075 |
| $R^2$ | 0.47 | 0.47 | 0.47 | 0.47 |
| Controls | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes |
| Forthcoming ratings | No | No | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of a design's maximum 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. The mean value of this variable in the sample is 0.32 (s.d. 0.27). Columns (2) and (4) control for time of submission and number of previous designs entered by the player and her competitors. Columns (3) and (4) additionally control for the best $f$orthcoming rating on the player's not-yet-rated designs. 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. SEs clustered by player in parentheses.

Table I.8: Similarity to player's best previously-rated designs & intra-batch similarity

| | Designs | | Batches (uwtd.) | | Batches (wtd.) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Player's prior best==5 | 0.351*** | 0.362*** | 0.214 | 0.238 | 0.249 | 0.285 |
| | (0.096) | (0.102) | (0.311) | (0.304) | (0.304) | (0.296) |
| * 1+ competing 5-stars | -0.204*** | -0.208*** | -0.302* | -0.305* | -0.300* | -0.295* |
| | (0.070) | (0.071) | (0.163) | (0.162) | (0.170) | (0.168) |
| * prize value ($100s) | -0.013 | -0.018 | 0.016 | 0.015 | 0.010 | 0.009 |
| | (0.031) | (0.033) | (0.099) | (0.097) | (0.095) | (0.093) |
| Player's prior best==4 | 0.119*** | 0.116*** | 0.050 | 0.065* | 0.062* | 0.086** |
| | (0.031) | (0.032) | (0.032) | (0.037) | (0.032) | (0.038) |
| Player's prior best==3 | 0.060** | 0.056** | 0.053 | 0.062* | 0.051 | 0.065* |
| | (0.028) | (0.028) | (0.035) | (0.037) | (0.035) | (0.037) |
| Player's prior best==2 | 0.026 | 0.024 | 0.018 | 0.027 | 0.006 | 0.018 |
| | (0.030) | (0.030) | (0.050) | (0.051) | (0.047) | (0.047) |
| 1+ competing 5-stars | -0.000 | 0.001 | 0.019 | 0.027 | 0.024 | 0.027 |
| | (0.022) | (0.024) | (0.048) | (0.049) | (0.052) | (0.054) |
| Days remaining | 0.000 | -0.007 | 0.001 | -0.008 | 0.000 | -0.005 |
| | (0.003) | (0.008) | (0.005) | (0.011) | (0.005) | (0.011) |
| Constant | 0.409** | 0.487*** | 0.383*** | 0.507*** | 0.386*** | 0.459*** |
| | (0.167) | (0.187) | (0.073) | (0.148) | (0.069) | (0.146) |
| N | 3871 | 3871 | 1987 | 1987 | 1987 | 1987 |
| $R^2$ | 0.53 | 0.53 | 0.57 | 0.57 | 0.58 | 0.58 |
| Controls | No | Yes | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations in Columns (1) and (2) are designs, and 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). Observations in Columns (3) to (6) 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), and 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 0.48 (s.d. 0.32). Columns (5) and (6) weight the batch regressions by batch size. All columns control for the time of submission 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. SEs clustered by player in parentheses.

Table I.9: Change in similarity to player's best previously-rated designs

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Δ(Player's best==5) | 0.878*** | 0.928*** | 0.914*** | 0.885*** | 0.929*** | 0.924*** |
| | (0.170) | (0.203) | (0.205) | (0.171) | (0.202) | (0.205) |
| * 1+ competing 5-stars | -0.411*** | -0.419*** | -0.427*** | -0.414*** | -0.418*** | -0.429*** |
| | (0.125) | (0.144) | (0.152) | (0.125) | (0.144) | (0.152) |
| * prize value ($100s) | -0.095** | -0.114** | -0.108** | -0.096** | -0.114** | -0.110** |
| | (0.039) | (0.049) | (0.047) | (0.040) | (0.049) | (0.048) |
| Δ(Player's best==4) | 0.281*** | 0.268*** | 0.276*** | 0.283*** | 0.270*** | 0.279*** |
| | (0.065) | (0.073) | (0.079) | (0.065) | (0.073) | (0.079) |
| Δ(Player's best==3) | 0.150*** | 0.135** | 0.137** | 0.151*** | 0.136** | 0.138** |
| | (0.058) | (0.065) | (0.069) | (0.058) | (0.065) | (0.069) |
| Δ(Player's best==2) | 0.082* | 0.064 | 0.059 | 0.082* | 0.063 | 0.059 |
| | (0.046) | (0.052) | (0.056) | (0.046) | (0.053) | (0.057) |
| 1+ competing 5-stars | -0.004 | -0.003 | 0.003 | -0.002 | -0.004 | 0.003 |
| | (0.014) | (0.014) | (0.024) | (0.015) | (0.014) | (0.026) |
| Days remaining | -0.001 | -0.001 | -0.001 | -0.005 | -0.001 | -0.004 |
| | (0.002) | (0.002) | (0.003) | (0.004) | (0.003) | (0.007) |
| Constant | -0.006 | -0.006 | 0.037 | 0.025 | -0.015 | 0.060 |
| | (0.009) | (0.008) | (0.066) | (0.040) | (0.037) | (0.077) |
| N | 2694 | 2694 | 2694 | 2694 | 2694 | 2694 |
| $R^2$ | 0.05 | 0.11 | 0.14 | 0.05 | 0.11 | 0.14 |
| Controls | No | No | No | Yes | Yes | Yes |
| Contest FEs | Yes | No | Yes | Yes | No | Yes |
| Player FEs | No | Yes | Yes | No | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of the *c*hange 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 experimenting to copying (and a value of -1, the converse). The mean value of this variable in the sample is -0.00 (s.d. 0.23). Columns (4) to (6) control for time of submission and number of previous designs entered by the player and 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. SEs clustered by player in parentheses.

Table I.10: 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 | 0.227 | -0.036 | -0.051 | -0.128 |
| | (0.300) | (0.225) | (0.244) | (0.154) |
| * 1+ competing 5-stars | -0.223 | -0.041 | -0.052 | 0.012 |
| | (0.176) | (0.114) | (0.132) | (0.099) |
| * prize value ($100s) | -0.029 | 0.005 | 0.010 | 0.028 |
| | (0.060) | (0.053) | (0.056) | (0.042) |
| Player's best forthcoming rating==4 | 0.137* | 0.006 | 0.003 | 0.000 |
| | (0.072) | (0.068) | (0.067) | (0.061) |
| Player's best forthcoming rating==3 | 0.146*** | -0.017 | -0.015 | -0.015 |
| | (0.054) | (0.100) | (0.098) | (0.097) |
| Player's best forthcoming rating==2 | 0.072 | -0.181* | -0.174* | -0.153 |
| | (0.051) | (0.101) | (0.098) | (0.094) |
| One or more competing 5-stars | -0.079 | 0.002 | 0.005 | -0.002 |
| | (0.102) | (0.116) | (0.122) | (0.123) |
| Days remaining | 0.011 | -0.038 | -0.038 | -0.040 |
| | (0.027) | (0.057) | (0.056) | (0.061) |
| Constant | -0.123 | 0.387 | 0.657 | 0.265 |
| | (0.185) | (0.464) | (0.495) | (0.489) |
| N | 1147 | 577 | 577 | 577 |
| $R^2$ | 0.68 | 0.83 | 0.83 | 0.67 |
| Controls | Yes | Yes | Yes | Yes |
| Contest FEs | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes |

Notes: Table provides a test of the effects of not-yet-available feedback on players' experimentation. Observations are designs. Dependent variable in Columns (1) to (3) is a continuous measure of a design's similarity to the best designs that the player has previously entered and has yet to *b*ut 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 in the sample is 0.26 (s.d. 0.25). If players depend on sponsors' ratings for signals of quality, then forthcoming ratings should have no effect on current experimentation. The results of Column (1) suggest this may not be the case; however, similarity to an unrated design may actually be the result of both these designs being tweaks on a third design. To account for this possibility, Column (2) controls for the given design's similarity to the best previously-rated design, the best not-yet-rated design's similarity to the best previously-rated design, and their interaction. Column (3) allows these controls to vary by the best 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 placebo that is not explained by jointly-occurring imitation of a third design. All columns control for time of submission 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. SEs clustered by player in parentheses.

Table I.11: Similarity to competitors' best previously-rated designs

| | (1) | (2) | | (3) | (4) |
|---|---|---|---|---|---|
| Competing best==5 | -0.029* | -0.003 | Δ(Competing best==5) | -0.002 | 0.000 |
| | -0.016 | -0.017 | | -0.046 | (0.046) |
| * 1+ own 5-stars | -0.011 | -0.011 | * 1+ own 5-stars | 0.027 | 0.027 |
| | -0.008 | -0.008 | | -0.027 | (0.027) |
| * prize value ($100s) | -0.005** | -0.017*** | * prize value ($100s) | -0.008 | -0.009 |
| | -0.003 | -0.003 | | -0.009 | (0.009) |
| Competing best==4 | 0.001 | 0.006 | Δ(Competing best==4) | 0.035 | 0.035 |
| | -0.013 | -0.013 | | -0.033 | (0.033) |
| Competing best==3 | 0.016 | 0.023* | Δ(Competing best==3) | 0.04 | 0.041 |
| | -0.012 | -0.012 | | -0.032 | (0.032) |
| Competing best==2 | 0.013 | 0.013 | Δ(Competing best==2) | 0.049 | 0.050 |
| | -0.014 | -0.014 | | -0.034 | (0.034) |
| One or more own 5-stars | 0.008 | 0.009 | One or more own 5-stars | 0.007 | 0.010 |
| | -0.027 | -0.028 | | -0.006 | (0.006) |
| Days remaining | -0.007*** | 0.002 | Days remaining | -0.001 | 0.000 |
| | -0.001 | -0.001 | | -0.001 | (0.002) |
| Constant | 0.059 | -0.002 | Constant | 0.006 | 0.004 |
| | -0.061 | -0.063 | | -0.022 | (0.029) |
| N | 9709 | 9709 | N | 6065 | 6065 |
| $R^2$ | 0.43 | 0.44 | $R^2$ | 0.11 | 0.11 |
| Controls | No | Yes | Controls | No | Yes |
| Contest FEs | Yes | Yes | Contest FEs | Yes | Yes |
| Player FEs | Yes | Yes | Player FEs | Yes | Yes |

Notes: Table provides a test of players' ability to discern the quality of, and then imitate, competing designs. Observations are designs. Dependent variable in Columns (1) and (2) is a continuous measure of the design's similarity to the highest-rated preceding entries by *o*ther players, taking values in [0,1], where a value of 1 indicates that the design is identical to another. The mean value in the sample is 0.14 (s.d. 0.10). Dependent variable in Columns (3) and (4) is a continuous measure of the *c*hange in designs' similarity to the highest- rated preceding entries by other players, taking values in [-1,1], where a value of 0 indicates that the player's current design is equally similar to the best competing design as was her previous design, and a value of 1 indicates that the player transitioned fully from experimenting to copying (and a value of -1, the converse). The mean value of this variable in the sample is 0.00 (s.d. 0.09). In general, players are provided only the distribution of ratings on competing designs; ratings of specific competing designs are not observed. Results in this table test whether players can nevertheless identify and imitate leading competition. Columns (2) and (4) control for time of submission 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. Robust SEs in parentheses.

Table I.12: Abandonment after a player's first rating, as function of rating

| | Dependent variable: Abandon after first rating | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linear | Linear | Linear | Logit |
| Player's first rating==5 | -0.352*** | -0.298*** | -0.481*** | -1.794*** |
| | (0.093) | (0.101) | (0.123) | (0.457) |
| * competing 5s | 0.038** | 0.024 | 0.076* | -0.044 |
| | (0.019) | (0.032) | (0.042) | (0.200) |
| Player's first rating==4 | -0.392*** | -0.386*** | -0.500*** | -1.945*** |
| | (0.050) | (0.068) | (0.083) | (0.267) |
| * competing 5s | 0.020 | 0.035 | 0.056* | -0.119 |
| | (0.015) | (0.025) | (0.031) | (0.176) |
| Player's first rating==3 | -0.302*** | -0.286*** | -0.374*** | -1.467*** |
| | (0.039) | (0.057) | (0.063) | (0.212) |
| * competing 5s | 0.013 | -0.000 | 0.028 | -0.184 |
| | (0.013) | (0.025) | (0.030) | (0.172) |
| Player's first rating==2 | -0.082** | -0.068 | -0.135** | -0.390* |
| | (0.037) | (0.055) | (0.062) | (0.206) |
| * competing 5s | -0.016 | -0.007 | 0.006 | -0.328* |
| | (0.012) | (0.025) | (0.030) | (0.170) |
| Competitors' prior best==5 | -0.040 | 0.030 | -0.025 | -0.453 |
| | (0.082) | (0.109) | (0.130) | (0.420) |
| Competitors' prior best==4 | 0.001 | 0.080 | 0.052 | -0.022 |
| | (0.063) | (0.085) | (0.091) | (0.307) |
| Competitors' prior best==3 | -0.079 | -0.002 | -0.025 | -0.376 |
| | (0.064) | (0.093) | (0.096) | (0.321) |
| Competing 5-star designs | 0.031** | 0.021 | 0.007 | 0.446** |
| | (0.013) | (0.024) | (0.029) | (0.178) |
| Days remaining | -0.019*** | -0.017*** | -0.026*** | -0.098*** |
| | (0.006) | (0.006) | (0.010) | (0.030) |
| Constant | 0.899*** | 0.817*** | 0.904*** | 1.961*** |
| | (0.080) | (0.095) | (0.206) | (0.720) |
| N | 1673 | 1673 | 1673 | 1635 |
| $R^2$ | 0.20 | 0.57 | 0.65 | |
| Contest FEs | Yes | No | Yes | Yes |
| Player FEs | No | Yes | Yes | No |

Notes: Table shows the effect of a player's first rating in a contest, and the competition at that time, on the probability that she subsequently enters more designs. Observations are contest-players. Columns (1) to (3) estimate linear models with fixed effects; Column (4) estimates a logit model without player fixed effects, which could render the results inconsistent. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by player in parentheses.

Table I.13: Abandonment after a given design, as function of player's ratings and competition

| | Dependent variable: Abandon after given design | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linear | Linear | Linear | Logit |
| Player's prior best rating==5 | -0.017 | -0.182*** | -0.154*** | -0.068 |
| | (0.033) | (0.053) | (0.044) | (0.179) |
| * competing 5s | 0.014** | 0.030*** | 0.027*** | 0.068** |
| | (0.007) | (0.009) | (0.008) | (0.032) |
| Player's prior best rating==4 | -0.054*** | -0.104*** | -0.091*** | -0.212*** |
| | (0.016) | (0.022) | (0.023) | (0.082) |
| * competing 5s | 0.016*** | 0.024*** | 0.023*** | 0.091*** |
| | (0.005) | (0.008) | (0.007) | (0.026) |
| Player's prior best rating==3 | -0.025 | -0.020 | -0.006 | -0.051 |
| | (0.015) | (0.020) | (0.019) | (0.073) |
| * competing 5s | 0.015** | 0.024*** | 0.020** | 0.074*** |
| | (0.006) | (0.009) | (0.009) | (0.027) |
| Player's prior best rating==2 | -0.018 | 0.026 | 0.031 | -0.023 |
| | (0.023) | (0.027) | (0.028) | (0.110) |
| * competing 5s | 0.007 | 0.027* | 0.026* | 0.035 |
| | (0.010) | (0.014) | (0.014) | (0.045) |
| Player's prior best rating==1 | -0.024 | 0.063 | 0.058 | -0.056 |
| | (0.037) | (0.038) | (0.039) | (0.176) |
| * competing 5s | -0.024 | -0.012 | -0.017 | -0.109 |
| | (0.016) | (0.021) | (0.020) | (0.075) |
| Competitors' prior best==5 | 0.119*** | 0.085*** | 0.137*** | 0.574*** |
| | (0.022) | (0.021) | (0.024) | (0.109) |
| Competitors' prior best==4 | 0.050*** | 0.020 | 0.065*** | 0.245*** |
| | (0.017) | (0.016) | (0.017) | (0.083) |
| Competing 5-star designs | 0.004 | -0.015*** | -0.003 | 0.020 |
| | (0.004) | (0.004) | (0.005) | (0.020) |
| Days remaining | 0.004 | 0.007** | 0.009* | 0.028 |
| | (0.005) | (0.003) | (0.006) | (0.022) |
| Constant | 0.217*** | 0.073** | 0.061 | -1.229*** |
| | (0.044) | (0.036) | (0.102) | (0.336) |
| N | 11758 | 11758 | 11758 | 11758 |
| $R^2$ | 0.07 | 0.26 | 0.28 | |
| Controls | Yes | Yes | Yes | Yes |
| Contest FEs | Yes | No | Yes | Yes |
| Player FEs | No | Yes | Yes | No |

Notes: Table shows the effects of feedback and competition at the time a design is entered on the probability that a player subsequently enters more designs. Observations are designs. Columns (1) to (3) estimate linear models with fixed effects; Column (4) estimates a logit model without player fixed effects, which could render the results inconsistent. All columns control for time of submission and number of previous designs entered by the player and her competitors. Results are qualitatively similar under a proportional hazards model. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. SEs clustered by player in parentheses.

Table I.14: Normalized panelist ratings on tweaks vs. experimental designs

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

Notes: Table compares professional graphic designers' ratings on tweaks and experimental 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, s.d. of panelists' ratings on a given design. A design is classified as a tweak if its maximum similarity to any previous design by that player is greater than 0.7 and an experiment if it is less than 0.3. Standard errors in parentheses below each mean; results from a one-sided test of equality of means is provided to the right. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively. Similarity scores calculated using perceptual hash algorithm. Results are robust to both algorithms and alternative cutoffs for experimentation.

# Part II

# Feedback in Tournaments for Ideas: Trading Off Participation for Quality

# Chapter Abstract

Performance feedback is a common feature of competitive settings in which new products are developed. While the adverse effects of feedback on incentives have recently received attention in the economics literature, its countervailing effect on productivity is generally overlooked: feedback not only informs agents of their rank, but also helps them improve. This paper uses a sample of four thousand commercial logo design tournaments to show that feedback reduces participation but increases the quality of new submissions, with an ambiguous net effect on high-quality output. To evaluate this tradeoff, I develop a procedure to estimate players' effort costs and use the estimates to simulate counterfactuals under alternative feedback policies. The results suggest that feedback on net increases the number of high-quality ideas produced and can thus be a desirable mechanism for a principal seeking innovation.

Feedback and evaluation are pervasive practices in industries where ideas and new products are developed: workers pitch their ideas to managers and clients, who provide critiques and direction; prototypes are tested at R&D labs, with focus groups, and in public demonstrations; and customers are the ultimate arbiters of value. Despite its widespread use, relatively little is known about the impacts of performance feedback on innovation in competitive environments. While feedback is argued to be essential to improving innovation (Manso 2011) and assuring its quality, research on the effects of feedback in competitive settings (e.g., Ederer 2010 and others) suggests that it can damage incentives by revealing asymmetries between competitors.

This paper studies the tension between incentives and improvement in feedback practices, which is of fundamental importance to firms in creative industries, as well as other organizations interested in promoting innovation. This tension is intrinsic to the product development process, and is particularly stark in formal competitions such as innovation prizes (Taylor 1995, Che and Gale 2003, Terwiesch and Xu 2008), which are undergoing a renaissance across private, public, and non-profit sectors (Williams 2012).[1] Innovation inducement prizes have been used for centuries to provide incentives for third parties to solve vexing commercial or social problems, and despite the attention-grabbing sums of popular examples, the mechanism is much more widely used: the America COMPETES Reauthorization Act of 2010 gave U.S. Federal agencies broad authority to conduct public prize competitions for problems large and small, and Challenge.gov has since hosted at least 280 contests from nearly 50 governmental agencies, with prizes ranging from status only (non-monetary) to tens of millions of dollars (OSTP 2014). In many cases, the sponsor has better information on performance than participants or can compel interim disclosures of progress and must decide whether to make it known while the competition is underway.[2]

A similar tension is present in non-innovation organizational settings, where performance appraisals serve the dual purposes of employee development and evaluation for tournament-like promotion and retention (Beer 1981, 1987; DeVries et al. 1981)., but this literature is short on empirical evidence and, with the exception of Wirtz (2014), does not account for the effects of feedback on agents' productivity. Another subliterature (e.g., Choi 1991, Gill 2008, Rieck 2010, Bimpikis et al. 2014, Halac et al. 2015) studies disclosure policies specifically in patent races and innovation contests, but it too is exclusively theoretical, and efforts to-date have excluded the possibility of feedback-driven improvement. I seek to add to the literature on both dimensions.

In this paper, I use a sample of four thousand winner-take-all commercial logo design contests to study the effects of feedback on the quantity and quality of submissions. I first show that feedback causes players to advantageously select into or out of participating and improves the quality of their designs, but disclosure of intense competition discourages effort from even the top performers. A principal seeking a high-quality product design thus faces a tradeoff between participation and improvement in deciding whether to provide its agents with feedback. To better

[1] Popular examples range from the 1714 British Longitudinal Prize for a method of calculating longitude at sea to recent, multi-million dollar X-Prizes for suborbital spaceflight and lunar landing. See Morgan and Wang (2010) for additional examples, and Terwiesch and Ulrich (2009) for practitioner-oriented discussion.

[2] In practice, sponsoring organizations often do. For example, in the 2006 Netflix contest to develop an algorithm that predicts users' movie ratings, entries were immediately tested and scored, with the results posted to a public leaderboard. In the U.S. Defense Advanced Research Projects Agency's 2005 and 2007 prize competitions to develop autonomous vehicles, participants had to publicly compete in timed qualifying races before moving on to a final round. Though the best-known examples have large (million-dollar) stakes, interim scoring is also common in contests for smaller-scale problems with lower stakes, such as architecture competitions or coding contests (Boudreau et al. 2011, Boudreau et al. 2014).

understand this tradeoff, I estimate a structural model of the setting and use the results to simulate tournaments with alternative feedback mechanisms. The results suggest that provision of feedback on net modestly increases the number of high-quality designs generated, with this increase entirely attributable to improvement rather than advantageous selection – implying that feedback is a valuable tool for generating innovation even in the presence of competition.

The paper begins by developing a simple model of a winner-take-all innovation contest to clarify the forces at play. In this model, a principal seeks a new product design and solicits candidates through a tournament, awarding a prize to the best entry. Players take turns submitting ideas, each of which receives immediate, public feedback revealing its quality. Partial-equilibrium predictions echo previous theoretical findings for other settings, particularly those of Ederer (2010): revelation of agents' performance can be motivating for high-performers, but in general will tend to disincentivize effort by exposing leaders and laggards. Yet the ostensible detriment to participation of providing feedback could nevertheless potentially be offset by the quality improvements it generates. Within this framework, I characterize feedback as having two effects: a *selection effect*, which drives players with poor reviews or facing fierce competition to quit, and a *direction effect*, which guides continuing players towards making increasingly better designs.

The paper then transitions to an empirical study of 4,294 commercial logo design competitions from a popular online platform. In these contests, a firm solicits custom designs from freelance designers, who compete for a winner-take-all prize awarded to the preferred entry. The contests in this sample typically offer prizes of a few hundred dollars and attract around 35 players and 115 designs. An essential feature of the setting is that the sponsor can provide real-time feedback on players' submissions in the form of 1- to 5-star ratings, which allow players to evaluate the quality of their own work and the competition they face. The first signs of a tension between participation and the quality of new submissions are apparent from correlating feedback provision with contest outcomes: contests in which a higher fraction of designs are rated attract fewer players and designs, but are also more likely to see sponsors award a retractable prize.

Using data at the contest-player and design level, I first provide evidence of the hypothesized effects. To identify an effect of feedback on quality, I examine (i) ratings on players' second designs, as a function of whether their first design was rated in advance; and (ii) improvements between consecutive submissions by a given player, where new information is made available but the latent ratings history is unchanged. I find that feedback does indeed improve subsequent entries, especially when the previous design was poorly-rated: for a player whose first design is rated 1-star, the probability that she improves with her second design increases from 26 percent to 51 percent when that rating is observed in advance (for those with a first design rated 2-stars, it is 27 to 41 percent; for 3-stars: 17 to 26 percent; for 4-stars: 6 to 11 percent).[3]

For evidence of an effect on participation, I estimate the probability that a player continues investing in or abandons a contest after their first rating. The likelihood of continuation is increasing monotonically in the first rating, with high performers continuing at a 50 percent higher rate than low performers. Yet even high performers can be driven away if their probability of winning is revealed to be close to either zero or one. Though these effects are clearly present, the results do not reveal whether the alleged trade-off between participation and quality is real, or which effect dominates. Since this question cannot be directly answered with the data in hand, I turn to structural estimation and simulations for insight.

[3]The second empirical strategy yields similar results.

I develop a theoretically-motivated procedure to estimate the success function and then the key parameter of the model that is unobserved: the cost of effort. The estimated costs are permitted to vary by player and contest and are identified from players' first-order conditions. In effect, I calculate (i) the expected payoff to each player's final design in a contest and (ii) the expected benefit of an additional, unentered design, and argue that the cost must be bounded by these two quantities. Because the contests in the sample typically attract dozens of players and hundreds of designs, the difference in the gains to a player's $n$th design and $(n+1)$th design will generally be small, and the estimated bounds are thus tight.

I then use the estimates to simulate contests under policies that isolate selection and direction, to demonstrate their individual and combined effects. I also simulate contests in which no designs are rated, all designs are rated, and a random subset are rated (according to the frequency of feedback observed in the data), to see if occasional feedback can outperform the alternatives by providing direction while limiting the fallout from selection. Both sets of simulations offer insights.

I find that direction has a dominant effect on the number of high-quality designs. When feedback operates only through selection (without improving quality), the number of top-rated designs declines by roughly 15 percent relative to a baseline of no feedback: despite selecting for better players, enough of them are driven away so as to reduce high-quality submissions. When improvement is enabled and selection suppressed, the number of top-rated designs explodes, increasing by over 900 percent: in effect, players learn to make increasingly better designs, and because they are oblivious to intensifying competition, they continue participating even as the tournament becomes excessively competitive. With both channels active, they cancel each other out, resulting in a more modest but positive, 10 percent increase in the number of top-rated entries. Providing feedback to only a random subset of designs increases the number of top-rated submissions by 35 percent relative to no feedback – outperforming comprehensive feedback.

Two implications follow. First, despite prior findings, feedback can be quite valuable in competitive settings when it improves the quality or productivity of agents' effort. However, feedback that merely selects for high performers is unlikely to increase high-quality innovation if they cannot leverage that feedback to improve their work. The key to getting the most out of a feedback policy in this setting is thus to provide guidance while limiting attrition. The second implication is that persistence is substantially more important to successful innovation in this setting than talent or luck: less talented players who respond to feedback will eventually outperform more talented players who ignore this feedback or otherwise fail to improve.

The paper proceeds as follows. Section II.1 discusses the literature in more depth and presents the theory. Section II.2 introduces the empirical setting. Section II.3 provides reduced-form evidence of the effects of feedback on participation and improvement. Sections II.4 and II.5 develop the structural model and characterize the cost estimates. Section II.6 presents results of the simulations. Section II.7 concludes.

## II.1 Feedback in Innovation Contests

### Existing Literature on Feedback in Tournaments

Rank-order tournaments have been the subject of a considerable amount of research since the seminal contributions of Tullock (1980) and Lazear and Rosen (1981), and the framework has been used to characterize competition in a wide variety of settings, most often workplace promotion. Interim evaluation in dynamic tournaments is a recent addition to the literature, motivated by the observation that "between 74 and 89 percent of organizations have a formal appraisal and feedback system" (Ederer 2010) and examples of evaluation in other settings.

This literature generally predicts that feedback will cause relative low-performers to reduce their investment and can incentivize high-performers to exert more or less effort. Ederer (2010) provides a nuanced view of the problem, showing that feedback can be motivating for high-performers, who learn their high productivity, but disclosure of asymmetries will discourage effort from both players, implying a tradeoff between what he terms "motivation" and "evaluation" effects. Such a result follows naturally from research on the effects of asymmetries on effort in tournaments, which consistently finds that incentives of both favorites and underdogs are reduced by unbalanced competition (e.g., Baik 1994, Brown 2011). Though a similar sorting effect arises in the present paper, the existing literature restricts attention to two-player competition. When there are many high-performing contestants, feedback may dampen incentives even if they are equally capable, since it reveals a crowded competition where the returns to effort are near zero.

The empirical evidence is both scarcer and more varied. Ederer and Fehr (2009) conduct an experiment in which agents select efforts over two periods and find that second-period effort declines in the revealed difference in first-round output. Eriksson et al. (2009) conduct an experiment in which competitors earn points for solving math problems and find that maintaining a leaderboard does not have significant effects on total attempts but can drive poor performers to make more mistakes, perhaps from adopting risky strategies in trying to catch up. Azmat and Iriberri (2010) examine the effects of including relative performance information in high schoolers' report cards on their subsequent academic performance, which has a tournament-like flavor. The evidence suggests that this information has heterogeneous but on average large, positive effects on treated students' future grades, which the authors attribute to increased effort.

The economics literature has also studied the question of whether to disclose performance in dynamic patent races and innovation contests. Choi (1991), for example, models a patent race and shows that information on rivals' interim progress has offsetting effects: it exposes a discouraging technological gap, but also changes perceptions of the success rate of R&D, which can be encouraging for laggards. Bimpikis et al. (2014) find similar results for R&D contests. Continuing the trend, Rieck (2010) shows that enforced secrecy yields the highest expected innovation in R&D contests, since disclosure only serves to introduce asymmetry. Finally, Halac et al. (2015) argue that in a setting where innovation is binary and the principal seeks one success, the principal's optimal mechanism either has disclosure and a winner-take-all prize (awarded to the first success, as in a race) or no disclosure and prize-sharing (split among successes).[4]

However, in none of these examples is feedback used to improve the quality of innovation or success rates. When the goal is to realize high-quality innovation, feedback can help workers learn, re-optimize with new strategies, and ultimately improve their product. My contribution here is thus to bring attention to the value of feedback as guidance, as Wirtz (2014) does for organizational settings, while recognizing potential adverse effects on incentives.[5]

[4]This model is also interesting in that it blends the features of contests and patent races.

[5]As Manso (2011) shows, this type of guidance is essential to motivating innovation in single-agent settings. Wirtz (2014) offers a similar rationale for feedback in tournament settings. In both cases, the instructive role of feedback is to inform the agent whether to stick to her current technology or experiment with a new one.

### Theoretical Underpinnings

Suppose a risk-neutral principal seeks a new product design.[6] Because R&D is risky and designs are difficult to objectively value, the principal cannot contract directly on performance and instead sponsors a tournament to solicit prototypes from $J$ risk-neutral agents, who enter designs in turns. At a given turn, a player must choose whether to continue participating and if so, what idea to develop next, with each submission receiving immediate, public feedback. At the end of the tournament, the sponsor awards a winner-take-all prize $P$ to its preferred entry.

To hone intuition, suppose each player enters at most two designs. Let each design be characterized by the following latent value $\nu_{jt}$, perhaps sponsor-specific, that only the sponsor observes:

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

where $j$ indexes players and $t$ indexes designs. In this model, $\beta_{jt}$ represents the design's quality, which may not be known ex-ante and is revealed by the sponsor's feedback. The design's value to the sponsor, $\nu_{jt}$, is increasing and concave in its quality, and the design with the highest $\nu$ wins the contest. The $\varepsilon_{jt}$ term is a random shock, which can be interpreted as idiosyncracies in the sponsor's tastes at the time a winner is chosen. Player $j$'s probability of winning is then:

$$Pr\,(\text{player } j \text{ wins}) = \frac{\beta_{j1} + \beta_{j2}}{\beta_{j1} + \beta_{j2} + \mu_j} \tag{II-2}$$

where $\mu_j \equiv \sum_{k \neq j} (\beta_{k1} + \beta_{k2})$ is the competition that player $j$ faces in the contest. This success function is effectively a discrete choice probability and obtains directly from the primitives.

Further suppose designs can be good, with quality $\beta^H$, or bad, with quality $\beta^L$. Every player's first design is an i.i.d. random draw from a distribution $F_\beta(\cdot)$, which yields a good design with probability $q$ or a bad design with probability $1-q$, in expectation $E[\beta] = q\beta^H + (1-q)\beta^L$. Players who receive positive feedback and choose to enter a second will recycle the high-quality design (with a new draw of the luck term), while those with negative feedback will make a new draw from $F_\beta(\cdot)$ – such that feedback yields weak improvement in expectation. Absent feedback, players may mix but will not consistently recycle a good design or make a new draw after a bad design. Moreover, players without feedback are ostensibly symmetric, entering designs with expected quality $E[\beta]$, and a player's perceived win probability is simply her share of submissions.

The remainder of this section provides two illustrative results in partial equilibrium, which conveys the basic intuition. First, I show that asymmetries reduce players' incentives to enter a second design and argue that feedback will tend to exacerbate these asymmetries, resembling the evaluation effect in Ederer's (2010) setting. I then show that provided competition is sufficiently high, as it typically is in the data, players with better feedback have greater incentives to continue participating than those with worse feedback, similar to Ederer's (2010) motivation effect.[7]

**Proposition 1.** *The returns to a player's second design decline as the quality of her first design and the cumulative competition grow distant, approaching zero at the limit. Feedback that reveals these asymmetries will therefore discourage participation, relative to a state of ignorance.*

[6] Some primitives of the model, and portions of the associated text, are borrowed from Gross (2014).
[7] Proofs are provided in Appendix II.A.

**Proposition 2.** *Upon provision of feedback, provided competition is sufficiently high, players with better feedback have higher incentives to participate than those with lower feedback.*

Intuitively, we might expect that negative feedback induces quitting: players with poor feedback not only face an uphill battle, but they are also more likely to produce lower-quality designs. Conversely, distant favorites can be equally unmotivated to exert effort, as victory is nearly assured – though this scenario is less likely to occur in a large field of competitors. In settings with many players, a third result emerges: high-performers who face heavy competition will also have low incentives to participate, because competition flattens the payoff curve to the point where marginal returns fall below cost. While these results have abstracted from the effects of strategic interactions, more explicit consideration would only tend to reinforce them. Additional effort would then yield indirect benefits to the inframarginal player by discouraging effort from followers (Ederer 2010), but this benefit would dissipate when competition is already high, since future competition is already deterred. This is similar to the notion of $\varepsilon$-preemption described in the context of patent races by Fudenberg et al. (1983).

## II.2 Graphic Design Contests

I turn to a sample of 4,294 logo design contests from a widely-used online platform to study the effects of feedback on participation and the quality of innovation in tournament competition. This platform 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 mock-ups, 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 intangible 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 a community of freelance designers (players) in exchange for a fixed prize awarded to its favorite entry. The sponsor publishes a design brief describing its business, its customers, and what it likes and seeks to communicate with its logo; specifies the prize structure; sets a deadline for entries; and opens the contest to competition. While the contest is open, 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. 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.

Appendix II.B describes the data gathering and dataset construction procedures in detail. For each contest in the sample, I observe the design brief, which includes a project title and description, 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. Though multiple prizes are possible, the sample is restricted to contests with a single, winner-take-all prize. I further observe every design submission, the identity of the designer, his or her history on the platform, the time and order in 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 and across all publicly observed contests in other design categories dating back to the platform's creation. I use the precise timing information to reconstruct the state of a 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 prior feedback available (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.

### Characteristics of the Sample

Table II.1 provides descriptive statistics for the contests in the sample. The average contest in the sample lasts nine days, offers a $295 prize, and attracts 116 designs from 37 players; on average, 66 percent of these are rated, but only three designs receive the top rating. By default, the sponsor retains the option of not awarding the prize to any design if none are to its liking, but the sponsor can forgo this option and commit to awarding the prize when it creates the contest. Though only 23 percent of contests have a committed prize, 89 percent are awarded.

[Table II.1 about here]

The median player in the sample competed in seven contests in any design category on the platform in the four years between the platform's launch in 2008 and August 1, 2012, when the data collection ended. This distribution is heavily skewed, with most players entering only one or two contests and a few participating frequently over extended periods. Less than a quarter of players in the sample have ever won a contest. The average winnings per submitted design is $4.58.

Table II.2 provides the distribution of ratings on rated designs. Fifty-eight percent of the designs in the sample (285,082 out of 496,041) are rated, slightly more than the average fraction of designs rated in a contest. The median and modal rating on these designs is three stars. Only five percent of all rated designs in the data receive the top, 5-star rating, suggesting that sponsors reserve this top rating for their most preferred entries. Indeed, a disproportionate number (almost 40 percent) of all winning designs are rated 5 stars, and nearly 75 percent are rated four or more stars, suggesting that

[8]One of the threats to identification throughout the empirical section is that the effect 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 that I do not observe. This concern is substantially mitigated by records of written feedback that were made available for a subset of contests in the sample. In cases where it is observed, written feedback is only given to a small fraction of designs in a contest (on average, 12 percent), far less than are rated, and typically echoes the rating given, with statements such as "I really like this one" or "This is on the right track". This written feedback is also not disproportionately given to higher- or lower-rated designs: the frequency of each rating among designs receiving comments is approximately the same as in the data at large. Thus, although the written commentary does sometimes provide players with explicit suggestions or include expressions of (dis)taste for a particular element such as a color or font, the infrequency and irregularity with which it is provided suggests that it does not supersede the role of the 1- to 5-star ratings in practice.

these ratings convey substantial information about a design's quality and odds of success, though they do not perfectly predict them: one- or two-star designs are occasionally observed to win contests (0.4 percent and 1.0 percent of awarded contests in the sample, respectively), suggesting that an element of luck remains until the end. Explanations for why low-rated designs sometimes win, or more generally why five-star designs do not always win, include last-minute changes of heart or differences of opinion between the member of the sponsoring organization administering the ratings and the person or committee selecting the winner.

[Table II.2 about here]

### Correlations of Contest Characteristics with Outcomes

To shed light on how the sampled contests operate and how different levers affect contest outcomes, Table II.3 explores the relationship of various outcomes with prize value, feedback, and other characteristics. The specifications in columns (1) to (3) regress the number of players, designs, and designs per player on the prize value, contest duration, length of the design brief, number of materials provided to be included in the design, and fraction of designs rated, as well as the average cost of participating players, which is estimated in later sections. Most of these variables are fixed by the sponsor before the contest begins, and while the fraction of entries rated and players' costs are in part endogenously determined during the contest, in practice they largely reflect the sponsor's type (engaged or aloof) and the difficulty of the project.

[Table II.3 about here]

An extra \$100 in prize value on average attracts around an additional 15 players and 55 designs. The effects of feedback are equally powerful: relative to a sponsor who rates no designs, one who rates every design will typically attract 14 fewer players and 20 fewer designs. Other features have more modest but nevertheless significant effects: contests with longer durations and longer design briefs tend to attract more submissions, but those with more complex design briefs are also less likely to be awarded.

In Column (7), I model of the probability that a sponsor chooses to award an uncommitted prize, implying that the contest generated a design good enough *to be* awarded. Feedback dramatically increases the probability that the prize is awarded, suggesting that feedback is critical to the development of high-quality work. In light of the aforementioned evidence that feedback reduces participation, this result provides the first indication of a tension between attracting *more* effort versus *higher-quality* effort. Contests with larger prizes, shorter design briefs, and lower costs are also more likely to be awarded, though the magnitude of these effects is considerably smaller than the effect of feedback: the effect of full feedback (relative to no feedback) on the probability the prize is awarded is nearly equal to that of a \$300 increase in the prize – more than doubling the average and median prizes in the sample.

## II.3 Reduced-form Evidence of Selection and Direction

### Effects of Feedback on Quality

In evaluating the effects of feedback on the quality of new submissions, a natural starting point is to examine the distribution of ratings on a player's second design, conditional on whether her first design was rated before the second was entered. As Figure II.2 shows, the overall mass of the ratings distribution shifts upwards when feedback is observed in advance – although this pattern could potentially be confounded by simultaneity, for example if sponsors are more likely to give timely feedback to high-performing players.

[Figure II.2 about here]

To alleviate this concern, I condition the comparison on the first design's rating. Figure II.3 shows the resulting conditional distributions:

[Figure II.3 about here]

Table II.4 provides the accompanying differences in means and shows that players improve at significantly higher rates when they observe feedback in advance. Among players with a 1-star design, 51 percent score a higher rating on their second entry when they observe feedback in advance, versus 26 percent among those who do not (Panel A). For players with a 2-star design, the percentages are 41 and 27 percent; for players with a 3-star design, 26 and 17 percent; and for players with a 4-star design, 11 and 6 percent, with all differences precisely estimated and significant the one percent level. Panel B shows the estimated effect on the second design's rating in levels, conditional on weak improvement.

[Table II.4 about here]

Though these specifications control for initial performance, they could nevertheless still be confounded if the players who are most likely to *improve* disproportionately wait for feedback before entering their second design. To resolve this issue, I turn to an alternative source of variation: pairs of consecutive submissions by a given player, in a given contest, between which the player's information set may change but her latent ratings history remains the same. This setting allows me to identify the effects of not only new information, but different levels of the feedback, while still conditioning on past performance. I estimate the following specification:

$$\Delta\text{Rating}_{ijk} \quad = \quad \beta_0 + \sum_{r=1}^{5} \beta_r \Delta\mathbb{1}(\text{Observed } r\text{-star ratings} > 0)_{ijk} + \varepsilon_{ijk}$$

conditioning on a player's latent performance being constant, where $\Delta\text{Rating}_{ijk}$ is the difference in the scores of successive designs by player $j$ in contest $k$, and the independent variables indicate the arrival of the player's first $r$-star rating in contest $k$ between designs $(i-1)$ and $i$.

Table II.5 provides the results. Panel A estimates a variant in which the outcome is an indicator for improvement ($\Delta\text{Rating}_{ijk} > 0$), while Panel B estimates the specification shown, moving the prior design's rating out of the first difference and into the constant to ease the interpretation. In both cases I further condition on the prior design's rating, to allow the effects on first differences to vary with the initial value.

[Table II.5 about here]

With few exceptions, revelation of feedback equal to or higher than the rating on the previous design leads to improvements (Panel A), while revelation of lower ratings generally has a precisely-estimated zero effect. Absent new information, new submissions tend to be of roughly the same quality as the prior submission (Panel B, constant), but feedback has large effects: players who enter 1- or 2-star designs and then receive a 5-star rating on an earlier submission tend to improve by over a full point more than they otherwise would. The magnitude of these effects expectedly declines the better the prior submission (i.e., from left to right across the table), reflecting the difficulty of improving from a high initial value.

### Effects of Feedback on Participation

Theory and intuition suggest that selection into and out of participating is a natural response to feedback. Though this effect can be discerned at various points in a contest, I again look to the first rating a player receives to identify the effect on participation. I focus on a player's first rating because it will typically be the first indication of the sponsor's preferences and ex-ante unpredictable, presenting an opportunity for clean identification. Appendix II.C provides estimates from analogous regressions on players' second ratings and finds similar results.

Table II.6 provides descriptive evidence of these effects. The table shows the distribution of the number of designs a player enters after receiving her first rating, conditional on whether that first rating is 1-star (left-most panel), or 4- to 5-stars (right-most panel). A majority (69.5 percent) of players whose first rating is the worst possible will subsequently drop out. In contrast, the majority (61.2 percent) of players who receive a high rating will subsequently enter additional designs.

[Table II.6 about here]

To formalize this result, I estimate the effect of each player's first rating and the competition perceived at the time it is made on the probability that the player subsequently abandons the contest, projecting an indicator for abandonment on indicators for each rating she may have received as well as indicators for the highest rating on competing designs:

$$\begin{aligned} Abandon_{jk} \quad &= \quad \beta_0 + \sum_{r=1}^{5} \beta_r \cdot \mathbb{1}(R_{jk} = r) + \sum_{r=1}^{5} \gamma_r \cdot \mathbb{1}(\bar{R}_{-jk} = r) \\ &\quad + \delta \cdot Timing_{jk} + X_{jk}\theta + \zeta_k + \varphi_j + \varepsilon_{jk} \end{aligned}$$

where $Abandon_{jk}$ indicates that player $j$ entered no designs in contest $k$ after her first rating; $R_{jk}$ is the player's first rating; $\bar{R}_{ijk}$ is the highest rating on any competing designs at that time; $Timing_{jk}$ measures the fraction of the contest elapsed at the time of that first rating; $X_{jk}$ is a vector of controls; and $\zeta_k$ and $\varphi_j$ are contest and player fixed effects, respectively. While it is likely the case that players with no observed activity after their first rating made a deliberate choice to stop participating, this measure cannot distinguish immediate abandonment from a "wait and see" strategy that ends in abandonment down the line. Since the result is the same, the distinction is immaterial for the purposes of this paper.

[Table II.7 about here]

Columns (1) to (3) estimate linear specifications with contest, player, and contest and player fixed effects, respectively. Linear specifications are used in order to control for these fixed effects (especially player fixed effects), which may not be estimated consistently in practice and could thus render the remaining estimates inconsistent in a binary outcome model. Column (4) estimates a logit model with only contest fixed effects. The linear model with two-way fixed effects (in Column 3) is the preferred specification.

The probability that a player enters more designs is monotonically increasing in that first rating: players with the most positive initial feedback are significantly more likely to remain active than those with poor initial feedback, and enter more designs at a precisely-estimated 50 percentage-point higher rate. High-rated competition also makes it more likely that a player abandons after her first rating. Altogether, these results establish that feedback leads to advantageous selection, with the high-rated players more likely to actively compete and low performers opting out after receiving low marks, but that by revealing high-rated competition, feedback can have the perverse consequence of driving everyone away.

To shed more light on the effects of asymmetries, I estimate a similar model replacing the indicators with a quadratic in a player's probability of winning upon receiving her first rating, which can be computed from the results of conditional logit estimation (described in detail in Section II.4 of this chapter). Table II.8 shows results from a similar arrangement of specifications: linear models with contest, player, and contest and player fixed effects in Columns (1) to (3), and a logit model with contest fixed effects in Column (4).

[Table II.8 about here]

The tendency to abandon is definitively convex in a player's probability of winning, reaching a minimum near a win probability of 0.5, and the estimates are statistically similar across all specifications – matching theoretical predictions that incentives for effort are greatest when agents are running even with their competition (Baik 1994). To visualize these results, Figure II.4 plots the predicted probability of abandonment against the win probability from a logit specification omitting the fixed effects.

[Figure II.4 about here]

### Results in Context

The collective evidence shows that feedback can have the desirable effects of improving the quality of future submissions and of nudging poor performers out of the contest, reducing wasteful effort and concentrating incentives for the remaining participants. But by revealing competition, feedback can reduce incentives for even the high-performers to participate, relative to the incentives in a state of ignorance. The principal thus faces a fundamental trade-off between participation and improvement. Given that sponsors who provide the most feedback are the most likely to award a retractable prize (recall Table II.3), it would seem that feedback has a large enough effect on quality to be desirable. Yet the reduced form is ultimately inconclusive and cannot discern how much of the increase in quality is due to better players or better designs. The distinction is not only important for understanding how to deliver feedback, but also revealing of the relative contributions of ability versus improvement to success in this setting.

If feedback were allowed to operate through only a selection or a direction channel, what would the competitive landscape look like, and how many high-quality designs would the principal receive? To answer this question, I develop a procedure to estimate the success function and players' costs of design, and use the estimates to simulate alternative feedback mechanisms. The simulations turn off the selection and direction channels one at a time in order to demonstrate the importance of each and their combined effect. I also compare outcomes when feedback is provided to zero, some, or all designs, since partial feedback may be sufficient to reap the benefits of direction while limiting the damage from selection.

## II.4 Structural Model

To disentangle the effects of selection and direction, I need an estimate of players' costs, which is the key parameter of tournament models not directly observed in the data. To obtain one, I develop a theoretically-motivated estimation procedure that bounds each player's cost in every contest in which she participates. The resulting estimates can then be used to simulate counterfactuals that decompose the effects of selection and direction on the quality of innovation in this setting.

The empirical model borrows ideas from the empirical auctions literature, which uses theoretical insights to estimate unobserved distributions of bidder values, and it is flexible in that costs are allowed to vary by contest and player, reflecting the fact that some contests are more demanding than others and that players have heterogeneous reservation wages. The main assumptions are (i) that each player has a constant cost in a given contest and (ii) that players compete until this cost exceeds the expected benefit. With a consistently estimated success function, a given player's cost in a given contest will be set-identified in a sample with *any* number of contests or players. The bounds of the identified set converge on a point as the number of players or designs in a contest grows large, irrespective of the number of contests in the sample.

The estimation proceeds in two steps. In the first step, I estimate a logistic success function that translates players' effort into their probability of winning. I then combine the success function with non-parametric frequencies of ratings on a player's next design, conditional on her prior history in the contest, to calculate the expected payoff to each player's last design in a contest and the "extra" design that the player chose not to enter. Under the assumption that the game ends in a complete information Nash equilibrium, these quantities place bounds on cost: a player's cost must be less than the expected benefit from her final design but greater than the expected benefit of an additional, unentered design. The logic behind this procedure is closely related to that of Haile and Tamer (2003), who use an analogous approach with drop-out bid levels to put bounds on the distribution of latent bidder values in symmetric English auctions.

Denote the rating of design $i$ from player $j$ in contest $k$ as $R_{ijk}$, and let $R_{ijk} = \emptyset$ when the sponsor declines to rate $ijk$. Ratings provide players with information on two unknowns: (i) the likelihood of a given design winning the contest, conditional on the competition, and (ii) how well her next design is likely to be rated. To make the intuition concrete, consider three cases:

1. $R_{ijk} = 5$. The design is outstanding and has a very high chance of winning the contest. The player has caught onto a theme that the sponsor likes. The player's subsequent designs are likely to be highly rated as well - though the marginal benefit of another submission fall dramatically, because the player already entered a design that is a strong contender. Any other five-star designs she enters will substantially cannibalize $ijk$'s odds.
2. $R_{ijk} = 1$. The design is not good and is unlikely to win the contest. The player hasn't yet figured out what the sponsor likes, and her next design will likely be poorly rated as well.
3. $R_{ijk} = \emptyset$. The player receives no feedback on her design. She has no new information on the sponsor's tastes, and the distribution of ratings on her next design is roughly unchanged.

In the model below, I formalize this intuition. The empirical model treats the design process as a series of experiments that adapts to feedback from the sponsor, as in the theoretical model of Section II.1 and in Gross (2014). Players use feedback to determine the probability that the rated design wins the contest, refine their experimentation, and set expectations over the ratings on any subsequent designs. Using the non-parametric distribution of ratings on a design in conjunction with a conditional logit model translating those outcomes into contest success, players then (1) calculate the expected benefit of another design, (2) compare it to their contest-, player-specific cost, and (3) participate until the costs exceed the benefits. The cost of design is set-identified from this stopping choice. I take the midpoint of the set as a point estimate.

The estimated cost will be the cost of making a single design and will be that which rationalizes the stopping choices observed in the data. I assume that this cost is constant for each player throughout a given contest. Design costs primarily reflect the opportunity cost of the time and resources a player expends in the activity of designing a logo. They thus reflect differences in reservations wages, which figure prominently when players can enter from anywhere in the world. But the estimated costs may also reflect (and net out) any unobserved payoffs in the form of learning, practice, and portfolio-building, all of which motivate players' participation. There may also be an unobserved payoff that accrues to the winner, such as a new client relationship; the expected value of this benefit will be captured in the estimates as well. Finally, the estimates will also reflect any level bias that a player has over the process determining her probability of winning. In effect, I will be measuring the costs that players behave as if they face.[9]

## Details: Framework and estimation

### Estimating the success function

Let $i$ index submissions, $j$ index players, and $k$ index contests. Suppose every contest $k$ has $J_k > 0$ risk-neutral players, and every player $j$ in contest $k$ makes $I_{jk} > 0$ submissions. Let $I_k = \sum_{j \in J_k} I_{jk}$ be the total number of designs in contest $k$. Players in contest $k$ compete for a prize $P_k$.

[9] For example, if a player has a high reservation wage, has a low unobserved payoff, or underestimates her chances of winning, she will exert limited effort and be perceived to have a high cost of design.

As in the theoretical model, I assume the sponsor awards the prize to its preferred design. Formally, let $\nu_{ijk}$ be the latent value of design $ijk$ to the sponsor of contest $k$, and suppose that this value is a function of the design's rating and an i.i.d. Type-I E.V. error. With six possibilities for ratings, a design's value can be written as the sum of fixed effects for each rating and an error term:

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

This specification is closely related to the theoretical model in equation (II-1), with the main difference being a restricted, discrete domain for feedback. The error term represents unpredictable variation in the sponsor's preferences and explains why 5-star designs do not always win. While the number and content of designs in the sponsor's choice set varies between contests, designs always share a common attribute in their rating, which is assumed to fully characterize the predictable component of a design's quality, including any written commentary provided to players but not observed in the dataset. The choice set is assumed to satisfy I.I.A.; in principle, adding a design of any rating to a given contest would reduce competing designs' chances of winning proportionally.[10] For contests with an uncommitted prize, the choice set includes an outside option of not awarding the prize, whose value is normalized to zero.

Under the model in equation (II-3), player $j$'s probability of winning is:

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

This success function can be estimated as a conditional logit model (McFadden 1974) using the win-lose outcome of every design in the sample. Results are provided in Appendix Table II.C.4, from which several patterns emerge. First, the value of a design is monotonically increasing in its rating, with only a 5-star rating being on average preferred to the outside option, and the fixed effects are precisely estimated. To produce the same change in the success function generated by a 5-star design, a player would need 12 4-star designs, 137 3-star designs, or nearly 2,000 1-star designs – so competition effectively comes from the top. As a measure of fit, the predicted odds-on favorite wins almost half of all contests in the sample. These results demonstrate that this simple model fits the data quite well and in an intuitive way, suggesting that ratings provide considerable information about a player's probability of winning.

**Calculating the expected benefit from a design**

To compute the expected benefit to a given player of entering an additional design, I consider all of the ratings the design may receive (ratings of 1 to 5, or no rating), calculate the incremental change in the success function under each rating, and take the weighted average, weighting by the non-parametric probability of obtaining each rating conditional on a player's history in the given contest. For players without any ratings in a given contest, these non-parametric frequencies are allowed to vary by quartile of their historical average to account for ex-ante heterogeneity in ability. This approach flexibly incorporates both ability and the learning that occurs with feedback and is central to the creative process.

[10] I also test the I.I.A. assumption by removing subsets of designs from each contest and re-estimating the model. The results are statistically and quantitatively similar when the choice set is deliberately varied.

**Distribution of outcomes for each design**

Let $s_{ijk}$ be a state variable characterizing the *eventual* ratings on all of player $j$'s designs in contest $k$ made *prior to* her $i^{th}$ design. $s_{ijk}$ can be summarized as a six-dimensional vector:

$$s_{ijk} = \left[ \sum_{x<i} \mathbb{1}(R_{xjk} = \emptyset), \sum_{x<i} \mathbb{1}(R_{xjk} = 1), \ldots, \sum_{x<i} \mathbb{1}(R_{xjk} = 5) \right]$$

In practice, a player's earlier designs aren't always rated before she makes her next one. Since $s_{ijk}$ incorporates all information on $j$'s past designs that will ever be known, and we can think of it as the ratings history under omniscience. However, players' experimentation choices must be made on the basis of prior, *observed* ratings, which are typically incomplete. Let $\tilde{s}_{ijk}$ be the ratings on previous submissions that player $j$ observes at the time of her $i^{th}$ submission in contest $k$. Writing $\tilde{R}_{xjk}$, $x < i$ as the rating on submission $x < i$ observed by the player at the time of her $i^{th}$ submission, we can write the observable ratings history as:

$$\tilde{s}_{ijk} = \left[ \sum_{x<i} \mathbb{1}(\tilde{R}_{xjk} = \emptyset), \sum_{x<i} \mathbb{1}(\tilde{R}_{xjk} = 1), \ldots, \sum_{x<i} \mathbb{1}(\tilde{R}_{xjk} = 5) \right]$$

The sum of the entries in the vector $\tilde{s}_{ijk}$ will equal the sum of those in $s_{ijk}$, but the null rating count may be higher for $\tilde{s}_{ijk}$, since some of the submissions made prior to $i$ that will eventually be rated have not yet been given a rating at the time the player makes her $i^{th}$ submission.

With a sample of 496,041 submissions, we can estimate the non-parametric distribution of the rating on $j$'s $i^{th}$ submission conditional on her contemporaneously observable ratings history $\tilde{s}_{ijk}$:

$$\hat{f}(R_{ijk} = r|\tilde{s}_{ijk}) = \frac{\sum_{\ell \in I_k, k \in K} \mathbb{1}(\mathrm{R}_{\ell k} = r|\tilde{s}_{\ell k} = \tilde{s}_{ijk})}{\sum_{\ell \in I_k, k \in K} \mathbb{1}(\tilde{s}_{\ell k} = \tilde{s}_{ijk})} \xrightarrow{d} f(r|\tilde{s}_{ijk})$$

In words, the probability that player $j$'s $i^{th}$ design in contest $k$ is rated $r$, given an observable ratings history of $\tilde{s}_{ijk}$, can be estimated from the data as the fraction of all designs in the data made in state $\tilde{s}_{ijk}$ that received the rating $r$. With a small, discrete sample space, these probabilities are easily estimated without the kernel methods required for continuous distributions.

The distribution $\hat{f}(\cdot)$ nevertheless suffers a curse of dimensionality due to the large heterogeneity in ratings histories. To reduce the dimensionality, I re-define $s_{ijk}$ and $\tilde{s}_{ijk}$ to indicate *whether* a player has received each rating, as opposed to counts of each rating. This adjustment is designed to make the non-parametric estimation tractable (with $2^6 = 64$ cells) while retaining the most important information in the ratings history. Under this construction, $\tilde{s}_{ijk}$ can be re-defined as follows:

$$\tilde{s}_{ijk} = \left[ \left( \sum_{x<i} \mathbb{1}(\tilde{R}_{xjk} = \emptyset) > 0 \right), \left( \sum_{x<i} \mathbb{1}(\tilde{R}_{xjk} = 1) > 0 \right), \ldots, \left( \sum_{x<i} \mathbb{1}(\tilde{R}_{xjk} = 5) > 0 \right) \right]$$

Figure II.1 illustrates some examples of $\hat{f}(\cdot)$. The top panel shows the distribution on a player's first design in a contest, and the bottom panel shows the distribution on a player's second design conditional on the first design receiving (from left to right): 1 star, no rating, and 5 stars. The results are intuitive: Players with high ratings enter better designs, players with low ratings enter worse designs, and players with no feedback draw from approximately the same distribution with their second design as with their first.

[Figure II.1 about here]

Heterogeneity in ability likely exists even in the absence of feedback. To account for this heterogeneity, I model players with no feedback as drawing from ability-specific distributions. For these cases, I estimate these distributions conditional on her quartile for average ratings in previous contests, adding an additional category for players who have no ratings in previous contests. In effect, this allows players with a good track record to draw their first design from a higher-quality bucket. However, once players have feedback on designs in a given contest, the estimation only conditions on feedback received in that contest, and players' track record in earlier contests is no longer relevant to the estimation.

I assume the players know the distributions $f(\cdot|\tilde{s})$ (or can infer it by intuition, experience, and examination of past contests) and plug in their observable ratings history, $\tilde{s}_{ijk}$, when they do the implicit cost-benefit calculations and decide whether to continue participating.

**Expected benefit from an additional design**

The expected rating on an additional design $I_{jk}+1$, at the time player $j$ chooses to make it, is:

$$E\left[R_{I_{jk}+1,jk}\right] = \sum_{r} r \cdot f(r|\tilde{s}_{I_{jk}+1,jk}) \ ,$$

the weighted average of all possible ratings, weighted by the probability of each. The expected increase in the player's odds of winning from the additional design, in a guaranteed-prize contest and holding the competition constant, can similarly be written as follows:

$$\begin{aligned} E\left[\Delta Pr\left(j \text{ wins } k\right)\right] &= \sum_{r} \left(\Delta Pr(j \text{ wins } k)|R_{I_{jk}+1,jk} = r\right) \cdot f(r|\tilde{s}_{I_{jk}+1,jk}) \\ &= \sum_{r} \left( \frac{e^{\beta_r} + \sum_{i \in I_{jk}} e^{\psi_{ijk}}}{e^{\beta_r} + \sum_{i \in I_k} e^{\psi_{ik}}} - \frac{\sum_{i \in I_{jk}} e^{\psi_{ijk}}}{\sum_{i \in I_k} e^{\psi_{ik}}} \right) \cdot f(r|\tilde{s}_{I_{jk}+1,jk}) \end{aligned}$$

The first term is the probability of winning with an additional design rated $r$, while the second term is the probability of winning without it. Their difference is the increase in player $j$'s odds from that design, which is weighted by the probability of an $r$-star rating and summed to get its expected value. The expected benefit of the design is this quantity multiplied by the prize:

$$E\left[\text{MB}_{I_{jk}+1,jk}\right] = E\left[\Delta Pr(j \text{ wins } k)\right] \cdot P_k$$

**Estimating costs from stopping choices**

Having obtained a success function from the sponsor's choice problem, and derived a non-parametric procedure for predicting quality, estimation requires two final assumptions:

- Players exert effort if the expected benefit exceeds the cost
- Players do not exert effort if the cost exceeds the benefit

The appeal of these assumptions is self-evident: they impose minimal requirements on agents and are nearly axiomatic in economic modeling.[11] To make the logic concrete, consider the final moments of a single contest. If designs can be made and entered in an infinitesimal amount of time (an assumption which, while hyperbolic, is perhaps not a bad approximation), then the contest should end in Nash equilibrium: given her ratings history and the ratings on competing designs, no player wants to experiment with another design. I similarly assume that a player's final design is a best response to competitors' play.

The implication is that the expected benefit of each player's last design exceeds her cost, whereas the expected benefit of the next design does not. These conditions allow me to place bounds on costs. Given the depth of these contests, the incremental benefit of an additional design is usually small, and the estimated bounds will therefore tend to be tight. I estimate each player's contest-specific cost of design to be bounded below by the expected benefit of the "extra" design that she does not to make ($I_{jk} + 1$) and bounded above by the ex-ante expected benefit of her final design ($I_{jk}$), as follows:

$$C_{jk} \in \left[\ E[\mathrm{MB}_{I_{jk}+1,jk}]\ ,\ E[\mathrm{MB}_{I_{jk},jk}]\ \right]$$

**Bootstrapped standard errors**

As functions of the MLE parameters and non-parametric frequencies, the estimated bounds are themselves random, taking the distribution of the convolution of their components. The maximum likelihood estimates are known to be normally distributed. A player's predicted success function at a given vector of efforts is thus the ratio of a sum of log-normals over a sum of log-normals. This ratio is calculated for the "final" design, subtracted from a similar quantity at the "extra" design, multiplied by the non-parametric probability of a given rating, and summed over all possible ratings to obtain the bounds. Randomness therefore enters from two sources: the MLE parameters and the non-parametric frequencies.

I use a block-bootstrap to obtain standard errors. To do so, I subsample entire contests from the dataset with replacement, re-estimate the logit parameters and non-parametric frequencies within the subsample, and use these estimates to re-calculate bounds on cost for every contest-player in the original dataset. My baseline bootstrap consists of 200 replications. As Section II.5 shows, the bounds are estimated precisely, and the identified set for each contest-player's cost is narrow relative to the midpoint.

[11] Haile and Tamer (2003) use similar assumptions to motivate their method of estimating bounds on bidder values in symmetric English auctions, which provided the foundations for the procedure developed in this paper.

### Assumptions and Identification

Identification of players' design costs hinges on four assumptions:

1. Costs are constant for a given player in a given contest (i.e., linear in effort).
2. The latent quality of each design is linear in its rating and an i.i.d. logit error.
3. The players know the distribution of ratings on their next design conditional on past ratings, as well as the process generating the odds of each design winning.
4. Players experiment up to the point where $E[\text{MB}] \leq \text{MC}$.

The assumption of linear costs could be reinterpreted as an approximation rather than a true assumption per se. The second assumption implies that all available information about quality is captured by a design's rating, and the reason 5-star designs do not win every contest effectively boils down to luck – in practice, the sponsor may change its mind, or different people might provide feedback versus award winners. Although the third assumption can be debated, these distributions are both intuitive and available to any player that has competed in or browsed past contests. The fourth assumption derives from economic theory.

For the purposes of estimation, I further assume that:

5. Players stop competing at the time they enter their last design.
6. At the time of this final submission, players have foresight over the state of the competition they will face at the end of the contest.

The fifth assumption is supported by conversations with professional designers who have participated on this platform and allege that they often enter their final design knowing it is their final design and simply "hope for the best." If players regularly checked in on each contest and decided whether to continue competing, the true time of abandonment would be unobserved. The final assumption is necessary for the agents to be able to compute the success function. On average, the majority of players in a contest exit in the last quarter. Since the distribution of ratings in the contest is publicly available, I assume that players know or can forecast the competition they will face at the end of the contest.[12] Appendix II.D offers descriptive evidence that the information available midway through a contest is sufficient to project the state of competition at the end of the contest reasonably well, supporting an assumption of foresight.

## II.5 Cost Estimates

Table II.9 provides summary statistics on the estimated costs of all 160,059 contest-players in the sample, calculated as the midpoint of the bounds estimated by the procedure described above. The cost estimates range from near zero to as high as \$108.24, with an average of \$5.77 and a median of \$4.62. To put these numbers in context, recall that the average per-design winnings for players in the sample, taken over their complete history on the platform, is \$4.58 (the median is zero).

[12] These players have available contemporaneous activity in the given contest, historical evidence from past contests (competed or observed), and intuition to make an informed assessment of competition.

[Table II.9 about here]

Figure II.5 shows the estimated bounds on cost for every contest-player in the sample. The upper bound is denoted in blue, and the lower bound in green. The red line traces the midpoint of these bounds for each player, which is my preferred point estimate for cost. Though confidence bands are not shown, these bounds are precisely estimated: the standard errors on the bounds are generally around 4.1 percent of the value (median 3.3 percent, 90th percentile 5.9 percent). They are also very tight, with the width of the identified set on average being 3.2 percent of the midpoint (median 2.6 percent, 90th percentile 6.2 percent), further motivating the choice of the midpoint as the preferred point estimate for simulations.

[Figure II.5 about here]

Upon seeing these results, the foremost question is whether the estimated costs are plausible. The mean estimated cost is around the average per-design winnings, suggesting that the estimates have approximately the right magnitude. What of their variation? Closer inspection of contests in which players are estimated to have very high costs typically reveals why. For example, in one such contest, players were asked to pay close attention to an unusually long list of requirements and provide a detailed, written explanation or story accompanying each design; as a result, only 23 designs were entered, in spite of a prize in the *99th percentile* of all contests. This result is a natural consequence of the model: since the expected benefits to an additional design will be high when participation is low, *ceteris paribus*, costs must be high to rationalize the observed exit patterns. Case studies of other contests further validate the estimates.

While these estimates are inevitably approximations, their quality is further evidenced by the fact that contest and player fixed effects explain nearly all (over 77 percent) of the variation in log costs, which should be the case if costs are primarily determined by the requirements of the contest and the characteristics of the player, and less so by the match between contests and players. Most of this variation (68 percent) is explained by contest fixed effects alone; in other words, costs vary considerably more between contests than within them. A smaller fraction (17 percent) is explained by player fixed effects alone.

The foremost shortcoming of the cost estimates is the imprecision that results from the dimensionality reduction in the model. Recall that the empirical model has players experimenting and projecting design outcomes on the basis of coarse ratings histories, with 64 variants. The low dimensionality is needed to make the estimation tractable, but it also sometimes results in multiple players having the same histories at the time of their stopping decisions and thus being estimated to have the same cost. This result is a reminder that the estimates are approximations, though it is not necessarily a reason to view them with skepticism. Approximations will generally be sufficient to account for costs in regressions (Section II.3) or use them to simulate contests under counterfactual conditions (Section II.6).

### Estimated costs are not strictly mechanical

The nature of this procedure raises the question of whether the estimated costs are mechanical or substantive economic quantities. Recall that costs are estimated at the midpoint of the payoffs to a player's final design and an extra, unentered design ($P_k \cdot \Delta\Pr(\text{Win}|\text{final design})$ and $P_k \cdot \Delta\Pr(\text{Win}|\text{extra design})$). Holding fixed the gains to the player's win probability, a one percent increase in prize would mechanically generate a one percent increase in the costs that I estimate. However in practice, the increase in a player's probability of winning from entering an additional design will not be fixed: theory predicts that players will compete away the value of larger prizes. The probability gains available at the player's final or extra design will then be reduced, offsetting the mechanical effect of prize increases in the cost estimation.

To test for the presence of such an offset, Appendix Table II.C.5 regresses the log probability gains achieved by a player's final design on the log prize. I find that larger prizes indeed tend to be heavily competed away: when the prize increases by one percent, the probability gains of players' final submissions declines by 0.75 percent – such that the residual cost estimate increases by only 0.25 percent. Though a perfect offset would manifest as an elasticity of -1, it should not be expected if projects with larger prizes are also more difficult, as the aforementioned evidence suggests. The regression results in Section II.2 additionally show that costs relate to contest outcomes in expected ways, with sensible magnitude, reinforcing the evidence that the cost estimates are real economic quantities and not merely an artifact.

## II.6 Decomposing the Effects of Feedback

I use the structural estimates to simulate tournaments under alternative feedback institutions which can clarify the role of feedback in this setting and separate its effects on participation and quality. In Section II.6, I compare total entry and the number of high-quality designs submitted when no, some, or all entries receive feedback. In Section II.6, I enable selection and direction individually to compare their separate and combined effects relative to a baseline of no feedback.

I sample 100 contests from the data and simulate three-player, sequential-play tournaments. Limiting the field to a few players reduces the dimensionality of the game sufficiently to be able to backwards-induct the best responses of each player's competitors and allow players to internalize these best responses, imposing discipline on their choices. The disadvantage to this approach is that the simulated contests have substantially fewer players than those in the sample, limiting the role that selection can play on the extensive margin (number of players). However, because the intensive margin (number of designs) features prominently, the presence of selection effects will still be highly detectable, as the results demonstrate.

The procedure is as follows. For every simulation of a given contest, I first (randomly) select three players in the contest and fix the order of their moves. The first player always enters. Beginning with player two, I project the distribution of ratings on her next design conditional on her ratings history in the contest, as well as the distribution of ratings that player three and then player one might subsequently receive, should they choose to engage. I then (i) evaluate player one's decision to participate or abandon conditional on the preceding outcome for players two and three, under a presumption of no further moves; (ii) backwards-induct player three's decision to participate or abandon conditional on the preceding outcome for player two and the anticipated response of player one; and finally (iii) backwards-induct player two's decision to participate or abandon given the distribution of responses from his competitors, choosing the action with the higher expected payoff.[13] If applicable, I then draw a design (in the form of a rating) for the given player, proceed to the next player, and repeat until every player has exited.

As in the empirical model, feedback enters players' decision-making in two places: it determines the distribution of the rating on a player's next design and the value and shape of her success function. The direction effect relates to the former, while the selection effect is a convolution of the two: players choose to remain active or abandon on the basis of the projected benefit to continuation, which is a function of both the rating on the next design and the incremental increase in the win probability that it generates. The net effects of feedback will reflect selection, improvement, and their interaction, but the advantage to simulation is that it makes it possible to gauge their relative importance by allowing one to vary while holding the other fixed, which is a possibility not otherwise supported by the data.

### Feedback vs. No Feedback

The first set of simulations compares contest outcomes under zero, partial, and full feedback policies. In these simulations, I draw each new design as a rating, according to the empirical frequencies in the data, irrespective of whether feedback is provided. Every design will thus have a latent rating, and the variation between feedback policies is in the disclosure of this rating to the players, which is implemented as follows. In the full feedback simulation, ratings are immediately revealed. In the simulation with partial feedback, designs will randomly have their ratings concealed, according to the empirical frequency of unrated designs in the sample. In the simulation without feedback, all ratings are concealed.

As context for the comparisons below, Table II.10 shows the distribution (across contests) of total submissions under each feedback policy, after averaging across simulations for each contest. Absent feedback-induced knowledge of asymmetries, simulated participation can grow quite large, but the design counts in simulations with feedback are comparable to those in the observed contests: the average simulated contest attracts 102 designs under partial feedback and 75 designs with comprehensive feedback.

[Table II.10 about here]

Table II.11 shows the effects of these feedback policies on the number of players, designs, and designs of each rating. The table shows the mean percent change in these outcomes across all contests and simulations as the policy varies from zero to partial feedback (Column 1), partial to full feedback (Column 2), and zero to full feedback (Column 3), with standard errors below. Given a pool of only three players, the number of players entering is relatively unaffected by feedback, as they will generally all participate: the reduction in extensive entry is (a statistically significant) 1.4 percent. However, moving from none rated to all rated on average reduces the number of entries in the simulations by over 81 percent.

[13] This approach to simulation was taken due to the absence of a closed-form solution amenable to plug-in simulation. Note that foresight remains partially limited under this procedure, as the player two moves ahead is forecast to behave as if she were the marginal player, which is necessary to make the dimensionality of the simulation tractable. However, the results suggest that this procedure sufficiently internalizes competitive best-responses, as the simulated number of designs in each contest are comparable to counts observed in the sample.

[Table II.11 about here]

Drilling down, the simulations suggest that feedback increases the number of top-rated designs and reduces the number of designs of every other rating, in some cases dramatically: comprehensive feedback leads to a 10 percent increase in 5-star designs and 90 percent reductions in 1- and 2-star designs, relative to no feedback. However, as demonstrated in the reduced-form, feedback is a double-edged sword, since it reveals competition and can reduce all players' incentives to participate. Partial feedback may be the more effective policy if it can support improvement while limiting the revelation of intensified competition. The results of these simulations suggest this is precisely the case: occasional feedback yields greater increases in top-rated designs, and more designs of every rating, than comprehensive feedback.

### Selection vs. Direction

The second set of simulations performs the focal exercise, toggling selection and direction. To turn off selection, I allow players to see only the number of own and competing designs and not the ratings on these designs when choosing whether to continue participating or abandon. To turn off direction, I hide the ratings from each player when a new design is drawn.[14]

Table II.12 provides the results. The first and second columns show the selection and direction effects individually, and the third column shows their combined effect for comparison. When feedback enters players' participation choices but not their design choices, as in Column (1), the total number of designs declines by 15 percent, with a roughly constant effect across ratings. When feedback enters design choices but not participation choices (Column 2), the number of top-rated designs increases more than *nine-fold* relative to no feedback. In this scenario, feedback aids players in improving their work, and the shrouding of asymmetries causes players to continue competing much longer than they would with complete information. When feedback is allowed to enter both the participation and design choices (Column 3), these countervailing forces offset, resulting in a tempered but nevertheless significant, positive effect on high-quality submissions and a sharp reduction in low-quality submissions. Together, the results imply that the increase in high-quality innovation in this setting is entirely attributable to improvement.

[Table II.12 about here]

## II.7 Conclusion

This paper studies the effects of interim feedback on creative production in competitive environments. Although practitioners consider feedback to be an essential feature of product development, the interaction of its effects on quality and incentives presents an intricate problem for the mechanism designer: feedback provides meaningful information that both generates quality improvements and reveals asymmetries between competitors that dampen incentives for effort. To my knowledge, this tradeoff has not previously been studied in the economics literature.

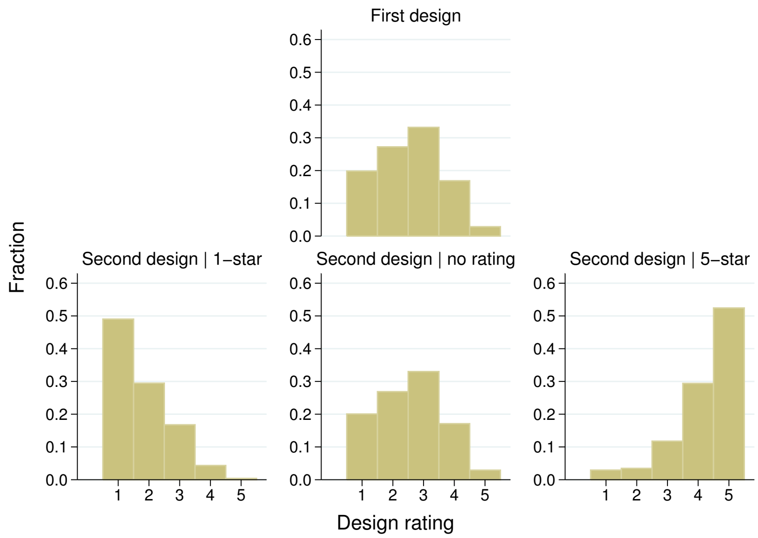
[14] These constraints could in concept arise in a real-world scenario in which the competitors were firms with managers and designers, and either (i) the designers could see the feedback but the managers deciding whether to continue competing were ignorant (i.e., a scenario without self-selection); or (ii) the managers deciding whether to continue competing do not pass along feedback they receive to the designers (scenario with no direction).

I use a sample of several thousand real-world commercial logo design tournaments to study this question. In these tournaments, a sponsor solicits custom designs from participants, provides feedback, and selects a winner after a pre-announced deadline. Reduced-form evidence corroborates the tension between improvement and participation: feedback both increases the quality of subsequent entries and motivates high performers, but the revelation of intense competition discourages continued participation. Since the net effect on high-quality submissions is ambiguous, I develop a structural procedure to estimate a function translating ratings into a probability of winning, as well as effort costs, and simulate tournaments under alternative feedback institutions. I compare outcomes when feedback is provided to all, some, or no designs, and when feedback is provided to all but the effect on quality or participation is muted.

The simulations are insightful. First, they suggest that complete feedback modestly increases the number of high-quality designs relative to no feedback. I also find that randomly administering feedback to only a subset of designs yields more high-quality submissions than complete feedback: partial feedback provides enough direction to yield improvements yet limits the amount of competition disclosed and the attrition this information generates. I then show that the positive effect of feedback on the quality of innovation is entirely the consequence of direction – implying that improvement is far more important to successful innovation in this setting than either talent or luck. It is also likely that the findings are externally valid: in many creative fields, including research, learning-by-doing and perseverance are crucial to achieving success.

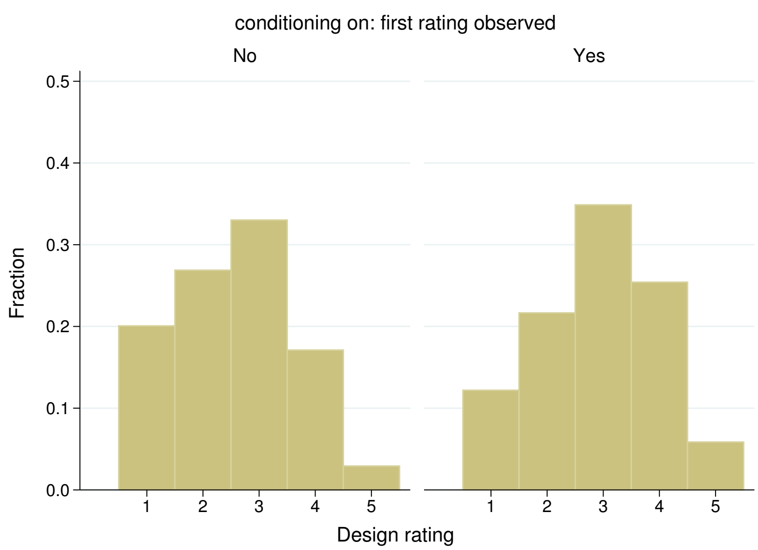
In light of the evidence in this paper, a natural opportunity for future research is to further explore the optimal frequency of feedback in organizations, comparing regular evaluation, no evaluation, and stochastic evaluation. An additional opportunity might be to compare a feedback mechanism to elimination, which is widely used in practice to pare down the competition and sharpen incentives for remaining finalists. Finally, this paper introduces a generalizable, structural approach to empirical research on tournaments, which has historically been constrained by a scarcity of large-sample data, and the methods developed in this paper can potentially be used or adapted to study tournaments in smaller samples.

Figure II.1: Distribution of ratings on 1st and 2nd designs



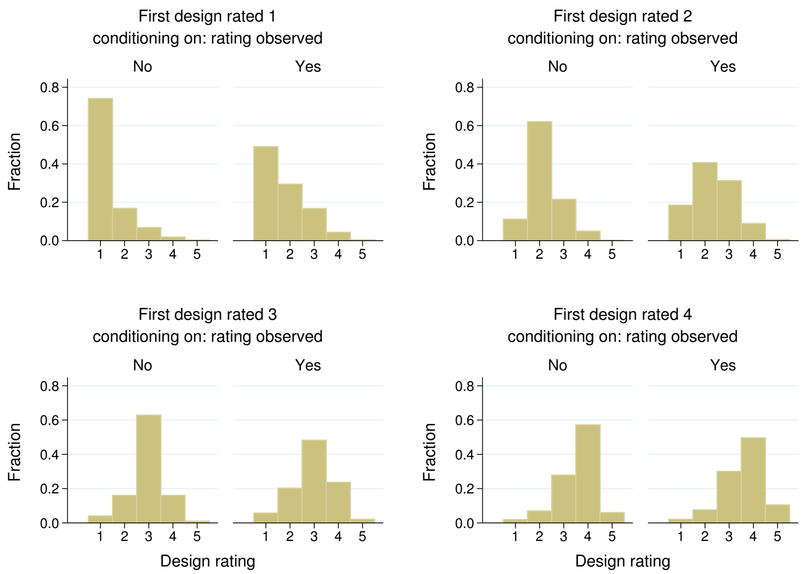

Notes: Figure shows distribution of ratings for: players' first design in a contest (top) and players' second design after receiving a 1-star rating on their first design (bottom left), after no rating (bottom center), and after a 5-star rating (bottom right).

Figure II.2: Distribution of ratings on 2nd design, by whether 1st design's rating observed



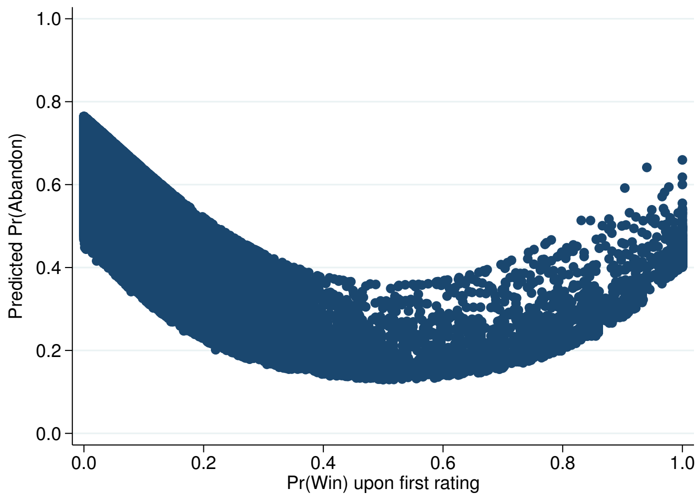

Notes: Figure reports the distribution of ratings on a player's second design in a contest conditional on whether that player's first design was rated prior to making the second one. The fraction rated 5 in the left panel is 0.03. The fraction rated 5 in the right panel is 0.06. The difference is significant with p=0.00.

Figure II.3: Distribution of ratings on 2nd design, by rating on 1st design and whether observed



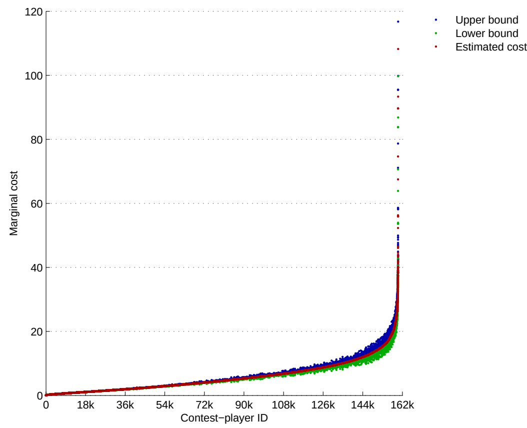

Notes: Figure shows the distribution of ratings on a player's second design in a contest, conditioning on (i) the rating they receive on their first design, and (ii) whether that rating was observed prior to entering their second design. In all subfigures, the fraction of players who score a better rating on their second design than on their first is higher when they observe feedback in advance (for players with a 1-star: 51 percent versus 26 percent; for those with a 2-star: 41 percent versus 27 percent; for those with a 3-star: 26 percent versus 17 percent; for those with a 4-star: 11 percent versus 6 percent). All differences are significant at the one percent level.

Figure II.4: Probability of abandoning after first rating




Notes: Figure shows the predicted probability that a player abandons the contest after her first rating, as a function of the contemporaneous win probability, estimated from a quadratic in Pr(Win) and controlling for timing. The win probability at which continued participation is most likely is 0.52.

Figure II.5: Estimated design costs, with bounds, in ascending order




Notes: Figure shows estimated costs for each contest-player in the sample. The upper bound is plotted in blue, and the lower bound in green. The red line traces the midpoint, which is the preferred point estimate. The figure arranges contest-players in increasing order of this midpoint.

Table II.1: Descriptive Statistics

| *Panel A. Characteristics of contests in the sample* | | | | | | |
|---|---|---|---|---|---|---|
| **Variable** | **N** | **Mean** | **SD** | **P25** | **P50** | **P75** |
| Contest length (days) | 4,294 | 9.15 | 3.72 | 7 | 7 | 13 |
| Prize value (US$) | 4,294 | 295.22 | 128.12 | 200 | 250 | 350 |
| No. of players | 4,294 | 37.28 | 25.35 | 23 | 31 | 43 |
| No. of designs | 4,294 | 115.52 | 94.82 | 65 | 92 | 134 |
| 5-star designs | 4,294 | 3.41 | 6.97 | 0 | 1 | 4 |
| 4-star designs | 4,294 | 13.84 | 17.89 | 3 | 9 | 19 |
| 3-star designs | 4,294 | 22.16 | 26.99 | 5 | 15 | 30 |
| 2-star designs | 4,294 | 16.04 | 23.36 | 2 | 8 | 21 |
| 1-star designs | 4,294 | 10.94 | 28.78 | 0 | 3 | 12 |
| Unrated designs | 4,294 | 49.14 | 63.36 | 10 | 34 | 65 |
| Number rated | 4,294 | 66.38 | 73.34 | 21 | 50 | 88 |
| Fraction rated | 4,294 | 0.56 | 0.33 | 0.3 | 0.6 | 0.9 |
| Prize committed | 4,294 | 0.23 | 0.42 | 0.0 | 0.0 | 0.0 |
| Prize awarded | 4,294 | 0.89 | 0.31 | 1.0 | 1.0 | 1.0 |
| *Panel B. Characteristics of contest-players in the sample* | | | | | | |
| **Variable** | **N** | **Mean** | **SD** | **P25** | **P50** | **P75** |
| No. of designs | 160,059 | 3.10 | 3.53 | 1.0 | 2.0 | 4.0 |
| Number rated | 160,059 | 1.78 | 2.95 | 0.0 | 1.0 | 2.0 |
| *Panel C. Characteristics of players in the sample* | | | | | | |
| **Variable** | **N** | **Mean** | **SD** | **P25** | **P50** | **P75** |
| No. of contests entered | 14,843 | 32.45 | 97.01 | 2 | 7 | 22 |
| Has ever won a contest | 14,843 | 0.19 | 0.39 | 0 | 0 | 0 |
| Winnings/contest | 14,843 | 14.68 | 75.27 | 0 | 0 | 0 |
| Winnings/submission | 14,843 | 4.58 | 24.76 | 0 | 0 | 0 |

Notes: Panel A 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. Panel B reports descriptives at the level of contest-players. Panel C reports descriptives across the entire sample for players who participated in at least one of the contests in the two-year sample used in the paper. These performance statistics calculations reflect a players' entire history on the platform through August 1, 2012 for all publicly visible contests (including design categories other than logos, which is the modal category).

Table II.2: Distribution of ratings (rated designs only)

| | **1-star** | **2-star** | **3-star** | **4-star** | **5-star** | **Total** |
|---|---|---|---|---|---|---|
| **Count** | 46,983 | 68,875 | 95,159 | 59,412 | 14623 | **285,052** |
| **Percent** | 16.48 | 24.16 | 33.38 | 20.84 | 5.13 | **100** |

Notes: Table tabulates rated designs by rating. 57.5 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 II.3: Correlations of contest outcomes with their characteristics

| | (1) Players | (2) Designs | (3) Designs/Player | (4) Awarded |
|---|---|---|---|---|
| Total Prize Value ($100s) | 14.828*** | 55.366*** | 0.124*** | 0.248*** |
| | (0.665) | (2.527) | (0.015) | (0.042) |
| Committed Value ($100s) | 1.860* | 5.584 | 0.008 | |
| | (1.118) | (4.386) | (0.025) | |
| Average Cost ($) | -1.790*** | -9.074*** | -0.088*** | -0.133*** |
| | (0.096) | (0.353) | (0.004) | (0.010) |
| Fraction Rated | -14.276*** | -20.056*** | 0.683*** | 0.691*** |
| | (0.812) | (2.855) | (0.040) | (0.106) |
| Contest Length | 0.340*** | 1.113*** | 0.003 | 0.007 |
| | (0.069) | (0.251) | (0.004) | (0.010) |
| Words in Desc. (100s) | 0.061 | 2.876*** | 0.059*** | -0.158*** |
| | (0.081) | (0.389) | (0.005) | (0.014) |
| Attached Materials | -0.878*** | -1.557** | 0.051*** | -0.011 |
| | (0.161) | (0.604) | (0.012) | (0.016) |
| Prize Committed | 1.076 | 2.909 | -0.023 | |
| | (3.290) | (12.867) | (0.085) | |
| Constant | 9.150*** | -4.962 | 2.488*** | 1.967*** |
| | (1.760) | (6.180) | (0.073) | (0.179) |
| N | 4294 | 4294 | 4294 | 3298 |
| $R^2$ | 0.63 | 0.65 | 0.31 | |

Notes: Table shows the estimated effect of contest attributes on overall participation and the probability that the prize is awarded. 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.

Table II.4: Effects of feedback: Improvement between first and second submissions

| ***P****anel A. Probability of improvement* | | | | |
|---|---|---|---|---|
| | When first design rated: | | | |
| | 1-star | 2-star | 3-star | 4-star |
| $\mathbb{1}$(First rating observed) | 0.251*** | 0.140*** | 0.087*** | 0.045*** |
| | (0.015) | (0.010) | (0.007) | (0.006) |
| Constant | 0.258*** | 0.267*** | 0.171*** | 0.060*** |
| | (0.006) | (0.005) | (0.004) | (0.003) |
| N | 8466 | 12653 | 16739 | 9192 |
| $R^2$ | 0.04 | 0.02 | 0.01 | 0.01 |

| ***P****anel B. Rating of second design* | | | | |
|---|---|---|---|---|
| | When first design rated: | | | |
| | 1-star | 2-star | 3-star | 4-star |
| $\mathbb{1}$(First rating observed) | 0.403*** | 0.260*** | 0.150*** | 0.079*** |
| | (0.026) | (0.016) | (0.010) | (0.010) |
| Constant | 1.373*** | 2.362*** | 3.226*** | 4.095*** |
| | (0.011) | (0.007) | (0.005) | (0.005) |
| N | 8466 | 11017 | 13072 | 5701 |
| $R^2$ | 0.04 | 0.03 | 0.02 | 0.01 |

Notes: Table shows the effects of observing feedback in advance of a player's second design in a contest on the probability that it is higher-rated than her first entry (Panel A) and on its rating, conditional on weakly improving (Panel B). *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors clustered by player in parentheses.

Table II.5: Effects of feedback: Improvement between any consecutive submissions

| ***P**anel A. Probability of improvement* | | | | |
|---|---|---|---|---|
| | When previous design rated: | | | |
| | 1-star | 2-star | 3-star | 4-star |
| $\Delta\mathbb{1}$(Obs. 5-star rating) | 0.184 | 0.389*** | 0.426*** | 0.274*** |
| | (0.155) | (0.114) | (0.047) | (0.030) |
| $\Delta\mathbb{1}$(Obs. 4-star rating) | 0.290*** | 0.358*** | 0.222*** | 0.001 |
| | (0.055) | (0.030) | (0.016) | (0.007) |
| $\Delta\mathbb{1}$(Obs. 3-star rating) | 0.258*** | 0.137*** | 0.030*** | -0.007 |
| | (0.036) | (0.020) | (0.008) | (0.012) |
| $\Delta\mathbb{1}$(Obs. 2-star rating) | 0.147*** | 0.119*** | -0.050*** | -0.013 |
| | (0.030) | (0.012) | (0.014) | (0.018) |
| $\Delta\mathbb{1}$(Obs. 1-star rating) | 0.195*** | -0.028 | -0.026 | -0.017 |
| | (0.015) | (0.026) | (0.024) | (0.029) |
| Constant | 0.253*** | 0.290*** | 0.215*** | 0.105*** |
| | (0.006) | (0.005) | (0.003) | (0.003) |
| N | 13012 | 19955 | 32637 | 21785 |
| $R^2$ | 0.04 | 0.02 | 0.01 | 0.01 |

| ***P**anel B. Rating of subsequent design* | | | | |
|---|---|---|---|---|
| | When previous design rated: | | | |
| | 1-star | 2-star | 3-star | 4-star |
| $\Delta\mathbb{1}$(Obs. 5-star rating) | 1.336** | 1.396*** | 0.695*** | 0.324*** |
| | (0.553) | (0.307) | (0.073) | (0.033) |
| $\Delta\mathbb{1}$(Obs. 4-star rating) | 0.999*** | 0.708*** | 0.260*** | 0.012 |
| | (0.155) | (0.065) | (0.020) | (0.010) |
| $\Delta\mathbb{1}$(Obs. 3-star rating) | 0.576*** | 0.194*** | 0.057*** | 0.010 |
| | (0.083) | (0.032) | (0.011) | (0.018) |
| $\Delta\mathbb{1}$(Obs. 2-star rating) | 0.212*** | 0.194*** | -0.020 | 0.009 |
| | (0.056) | (0.018) | (0.022) | (0.032) |
| $\Delta\mathbb{1}$(Obs. 1-star rating) | 0.303*** | 0.041 | -0.004 | -0.019 |
| | (0.027) | (0.044) | (0.037) | (0.049) |
| Constant | 1.373*** | 2.396*** | 3.282*** | 4.147*** |
| | (0.012) | (0.007) | (0.005) | (0.004) |
| N | 13012 | 17796 | 27025 | 15360 |
| $R^2$ | 0.05 | 0.04 | 0.02 | 0.01 |

Notes: Table shows the effects of newly-arrived feedback on the probability that a given design is higher-rated than that player's previous submission (Panel A) and on its rating, conditional on weakly improving (Panel B). These effects are identified by comparing the ratings on successive entries by a given player, in a given contest, where the player has the same latent ratings at the time of both entries but experiences change in her information set between the two as a result of newly-arrived feedback. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels. Standard errors clustered by player in parentheses.

Table II.6: Designs entered after a player's first rating

| **When first rating is 1 (out of 5)** | | | **When first rating is 4 or 5 (out of 5)** | | |
|---|---|---|---|---|---|
| **Designs after 1st rating** | **Freq.** | **Pct.** | **Designs after 1st rating** | **Freq.** | **Pct.** |
| 0 designs | 11,378 | 69.49 | 0 designs | 7,399 | 38.80 |
| 1 design | 1,696 | 10.36 | 1 design | 2,978 | 15.61 |
| 2 designs | 1,084 | 6.62 | 2 designs | 2,339 | 12.26 |
| 3 designs | 611 | 3.73 | 3 designs | 1,662 | 8.71 |
| 4 designs | 403 | 2.46 | 4 designs | 1,223 | 6.41 |
| 5 designs | 332 | 2.03 | 5 designs | 854 | 4.48 |
| 6 designs | 208 | 1.27 | 6 designs | 621 | 3.26 |
| 7 designs | 157 | 0.96 | 7 designs | 436 | 2.29 |
| 8 designs | 90 | 0.55 | 8 designs | 292 | 1.53 |
| 9 designs | 69 | 0.42 | 9 designs | 264 | 1.38 |
| 10+ designs | 346 | 2.11 | 10+ designs | 1,004 | 5.26 |
| **Total** | **16,374** | **100** | **Total** | **19,072** | **100** |

Notes: Table reports the activity of players after receiving their first rating in a contest, by the value of that first rating. A total of 86,987 contest-players received first ratings. Of these: 16,374 were rated 1 star (18.8 percent); 22,596 were rated 2 stars (26.0 percent); 28,945 were rated 3 stars (33.3 percent); 16,233 were rated 4-star (18.7 percent); and 2,839 were rated 5-star (3.3 percent). The table illustrates that players are much more likely to continue participating in a contest after positive feedback.

Table II.7: Tendency to abandon after first rating, as function of rating

| | Dependent variable: Abandon after first rating | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linear | Linear | Linear | Logit |
| Player's first rating==5 | -0.444*** | -0.394*** | -0.485*** | -2.311*** |
| | (0.017) | (0.017) | (0.020) | (0.086) |
| Player's first rating==4 | -0.437*** | -0.385*** | -0.454*** | -2.269*** |
| | (0.010) | (0.010) | (0.012) | (0.055) |
| Player's first rating==3 | -0.280*** | -0.242*** | -0.290*** | -1.468*** |
| | (0.008) | (0.008) | (0.009) | (0.044) |
| Player's first rating==2 | -0.114*** | -0.097*** | -0.120*** | -0.620*** |
| | (0.007) | (0.007) | (0.008) | (0.038) |
| Competitors' prior best==5 | 0.037*** | 0.064*** | 0.058*** | 0.203*** |
| | (0.013) | (0.011) | (0.014) | (0.065) |
| Constant | 0.483*** | 0.465*** | 0.174 | -0.898 |
| | (0.026) | (0.020) | (0.106) | (1.130) |
| N | 48125 | 48125 | 48125 | 46935 |
| $R^2$ | 0.24 | 0.36 | 0.46 | |
| Contest FEs | Yes | No | Yes | Yes |
| Player FEs | No | Yes | Yes | No |

Notes: Table shows the effect of a player's first rating in a contest and the competition at that time on the probability that the player subsequently enters no more designs. Observations are contest-players. The dependent variable in all columns is an indicator for whether the player abandons after her first rating. Columns (1) to (3) estimate linear models with fixed effects; Column (4) estimates a logit model without player fixed effects, which may render the estimates inconsistent. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. All specifications control for time remaining, both in levels and as a percent of the contest duration. Standard errors clustered by player in parentheses.

Table II.8: Tendency to abandon after first rating, as function of Pr(Win)

| | Dependent variable: Abandon after 1st rating | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linear | Linear | Linear | Logit |
| Pr(Win) | -1.643*** | -1.565*** | -1.644*** | -8.756*** |
| | (0.049) | (0.053) | (0.055) | (0.280) |
| Pr(Win)$^2$ | 1.565*** | 1.478*** | 1.537*** | 8.284*** |
| | (0.055) | (0.061) | (0.061) | (0.306) |
| Constant | 0.398*** | 0.379*** | -0.052 | -1.622 |
| | (0.025) | (0.019) | (0.085) | (1.389) |
| N | 48125 | 48125 | 48125 | 46935 |
| $R^2$ | 0.20 | 0.34 | 0.43 | |
| Contest FEs | Yes | No | Yes | Yes |
| Player FEs | No | Yes | Yes | No |
| Minimizer | 0.52 | 0.53 | 0.53 | 0.53 |

Notes: Table shows the effect of a player's win probability after receiving her first rating on the probability that she subsequently enters no more designs. Observations are contest-players. The dependent variable in all columns is an indicator for whether the player abandons after her first rating. Columns (1) to (3) estimate linear models with fixed effects; Column (4) estimates a logit model without player fixed effects, which may render the estimates inconsistent. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. All specifications control for time remaining, both in levels and as a percent of the contest duration. Standard errors clustered by player in parentheses.

Table II.9: Summary statistics for estimated costs

| | **N** | **Mean** | **SD** | **Min** | **P25** | **P50** | **P75** | **Max** |
|---|---|---|---|---|---|---|---|---|
| Est. Cost | 160,059 | 5.77 | 4.76 | 0.04 | 2.15 | 4.62 | 8.06 | 108.24 |

Table II.10: Distribution of simulated contests' submission counts under each feedback policy

| Simulated policy | N | Mean | SD | P10 | P25 | P50 | P75 | P90 |
|---|---|---|---|---|---|---|---|---|
| No feedback | 100 | 486.29 | 571.65 | 60.5 | 120.86 | 263.43 | 681.71 | 1224.09 |
| Partial feedback | 100 | 101.72 | 112 | 19.15 | 29.99 | 61.26 | 134.04 | 243.98 |
| Full feedback | 100 | 75.24 | 83.13 | 12.71 | 22.75 | 48.07 | 97.67 | 187.14 |

Notes: Table shows the distribution (across contests) of design counts for each simulated feedback policy: (i) no designs receive ratings, (ii) a subset receive ratings, and (iii) all receive ratings. Design counts for each contest are first averaged across its 50 simulations.

Table II.11: Effects of Feedback on Outcomes of Simulated Tournaments

| | Percent change in outcome, when: | | | | | |
|---|---|---|---|---|---|---|
| **Outcome** | **Some rated, relative to none rated** | | **All rated, relative to some rated** | | **All rated, relative to none rated** | |
| Players | -0.7% | *** | -0.7% | *** | -1.4% | *** |
| | (0.1%) | | (0.1%) | | (0.1%) | |
| Designs | -75.1% | *** | -6.2% | *** | -81.3% | *** |
| | (0.2%) | | (0.2%) | | (0.2%) | |
| Num. 5-star | 35.3% | *** | -24.9% | *** | 10.4% | *** |
| | (1.9%) | | (1.9%) | | (1.5%) | |
| Num. 4-star | -53.5% | *** | -8.9% | *** | -62.4% | *** |
| | (0.6%) | | (0.7%) | | (0.6%) | |
| Num. 3-star | -74.6% | *** | -6.4% | *** | -81.1% | *** |
| | (0.3%) | | (0.4%) | | (0.3%) | |
| Num. 2-star | -83.3% | *** | -5.3% | *** | -88.6% | *** |
| | (0.3%) | | (0.3%) | | (0.2%) | |
| Num. 1-star | -88.3% | *** | -3.9% | *** | -92.2% | *** |
| | (0.2%) | | (0.3%) | | (0.2%) | |

Notes: This table illustrates the effect of feedback on principal outcomes in simulated design contests. One hundred contests were randomly selected from the sample and simulated 50 times with three randomly-selected players. Simulations were performed under three scenarios: (i) feedback not provided; (ii) feedback randomly provided to a subset of designs, drawn according to the frequencies in the data; and (iii) feedback provided to all designs. In all cases, feedback is made available immediately after the player submits the design. The table provides the average percent change in the given outcome relative to a baseline simulation with no feedback, averaged over all simulations of all contests. Feedback is seen to reduce participation and increase the frequency of high-quality designs. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors in parentheses.

Table II.12: Effects of Feedback on Outcomes of Simulated Tournaments

| | Percent change in outcome, when: | | | | | |
|---|---|---|---|---|---|---|
| **Outcome** | **Selection/no direction, rel. to no feedback** | | **Direction/no selection, rel. to no feedback** | | **Combined effects, rel. to no feedback** | |
| Players | -1.1% | *** | 0.0% | n.a. | -1.4% | *** |
| | (0.1%) | | (0.0%) | | (0.1%) | |
| Designs | -15.1% | *** | 0.0% | n.a. | -81.3% | *** |
| | (0.8%) | | (0.0%) | | (0.2%) | |
| Num. 5-star | -17.5% | *** | 941.7% | *** | 10.4% | *** |
| | (1.0%) | | (8.9%) | | (1.5%) | |
| Num. 4-star | -16.6% | *** | 98.7% | *** | -62.4% | *** |
| | (0.8%) | | (1.0%) | | (0.6%) | |
| Num. 3-star | -14.9% | *** | -19.6% | *** | -81.1% | *** |
| | (0.8%) | | (0.3%) | | (0.3%) | |
| Num. 2-star | -11.4% | *** | -54.0% | *** | -88.6% | *** |
| | (0.9%) | | (0.3%) | | (0.2%) | |
| Num. 1-star | -12.0% | *** | -66.4% | *** | -92.2% | *** |
| | (0.9%) | | (0.3%) | | (0.2%) | |

Notes: This table separates feedback's effects on quality and participation in the simulated contests. Effects are isolated by running simulations in which feedback is allowed to enter players' decisions to continue or drop out but not influence experiment outcomes (Column 1), and vice versa (Column 2). In effect, the simulations make feedback visible to players when deciding whether to continue or drop out but invisible/unavailable when they draw their next design (and vice versa). One hundred contests were randomly selected from the sample and simulated 50 times with three randomly-selected players. Simulations were performed under three scenarios: (i) feedback is available for the continuation decision but not experimentation; (ii) feedback is available for creative choices but not the continuation decision; and (iii) feedback is available for both choices. The table provides the average percent change in the given outcome relative to a baseline simulation with no feedback, averaged over all simulations of all contests. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Standard errors in parentheses.

# Part III

# Scale versus Scope in the Diffusion of New Technology

## Chapter Abstract

Using the farm tractor as a case study, I show that lags in technology diffusion arise along two distinct margins: scale and scope. Though tractors are now used in nearly every agricultural field operation and in the production of nearly all crops, they first developed with much more limited application, and early diffusion was accordingly limited in scope until tractor technology generalized. Other historically important innovations share similar histories. The results suggest that the key to understanding technology diffusion is not only in explaining the number of different users, but also in explaining the number of different *u*ses.

Technology diffusion is widely held to be a leading explanation for productivity growth and productivity differences across industries, firms, and geographic regions. For example, it is frequently argued that facilitating the diffusion of modern agricultural technologies to developing countries is a key to lifting incomes and breaking a cycle of poverty. Previous research on technology diffusion has focused heavily on explaining variation in the scale of diffusion for particular applications. Considerably less attention has been paid to determinants of its scope – the set of applications in which the focal technology is used at all – despite Griliches' (1957) contention that this extensive margin is one of the principal dimensions along which diffusion occurs.[1]

In this paper, I show that the key to understanding lags in the diffusion of the farm tractor and other historically important innovations lies not only in explaining the number of users, but also in explaining the number of *u*ses. Each of the examples in this paper – tractors and hybrid seed corn in agriculture, and steam and electric power in manufacturing – first developed for applications with exogenously high demand, and initial diffusion was accordingly limited in scope. Only later did these technologies become sufficiently general to be useful for other purposes. The history is consistent with economic theory, which suggests that R&D will naturally progress from specific- to general-purpose variants of an innovation, and that these technical advances will directly translate to increased scope for diffusion. Lags in diffusion will therefore often be the consequence of holdups and market failures in R&D that stymie the generalization of new technology.

To clarify the forces underlying changes in the scope of diffusion, I begin by developing a model of innovation with R&D in specific- versus general-purpose technological attributes.[2,3] The model intuitively predicts that product features will endogenously be developed in the order in which they are most valuable, implying that new technologies will first be invented for narrow, high-value applications and only later – if at all – generalize for broader use. Diffusion will thus tend to follow an S-shaped pattern not only within applications, but also across them. Complementarities between the given technology and other innovations can drive a wedge between the private and social returns to investing in a general-purpose variant, and inventing firms will therefore often be suboptimally incentivized to invest in expanding the scope of their technology.

The paper then transitions to an empirical study of tractor diffusion, followed by a shorter survey of hybrid corn, steam engines, and electric power. Though tractors are now used in nearly every agricultural field operation and in the production of nearly all crops, they first developed for much more limited applications of tillage and harvesting grain. Recent research has emphasized the role of factor price changes and quality improvements in explaining aggregate diffusion (Manuelli and Seshadri 2014), but the literature is missing a crucial part of the story: tractor quality historically varied as much or more across space as it did over time. Indeed, its significance today is the

[1] As Griliches (1957) shows, the diffusion of hybrid corn in each U.S. crop-reporting district was defined not only by the rate at which it proceeded, but also by when it began. Most research on diffusion focuses on the former, which has been attributed to heterogeneous costs/benefits (Duflo et al. 2008, Suri 2011), fixed costs of adopting an indivisible technology (David 1966, Olmstead 1975), and changes in relative factor prices (Manuelli and Seshadri 2014), as well as to suboptimal decision-making due to credit constraints (Clarke 1991), information spillovers (Conley and Udry 2010, Dupas 2014, Munshi 2004), and individual biases (Duflo et al. 2011). The onset of diffusion, on the other hand, is a consequence of what Griliches termed the "availability" problem: hybrid seed corn had to be adapted to growing conditions of different crop-reporting districts before it could be locally grown.

[2] The model builds on the theoretical framework developed by Bresnahan and Trajtenberg (1995) to characterize general-purpose technologies. Though this paper is not a study of general-purpose technologies, the framework is a useful starting point for thinking about the path of R&D and other issues at hand.

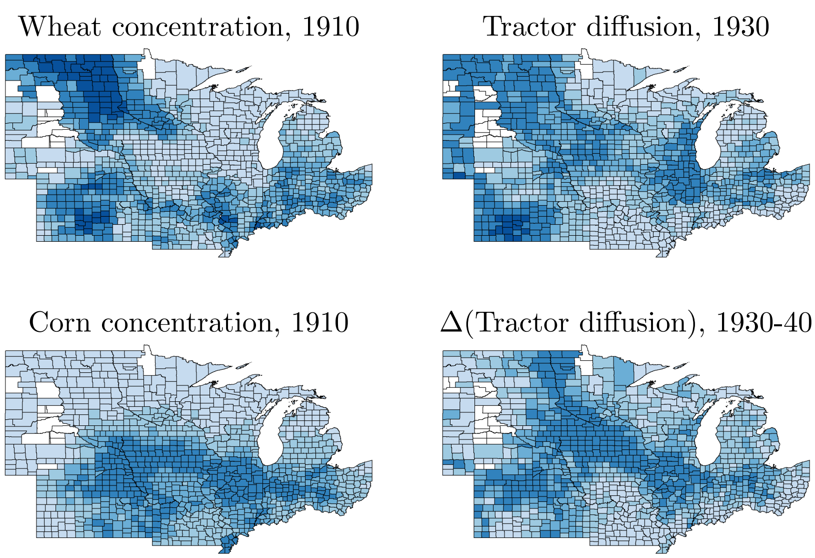
[3] The model also has analogy to Lazear's (2009) study of specific and general human capital.

consequence of not only its mechanical efficiency, but also its versatility as a source of mechanical power in agriculture. The other examples studied here share a similar history.

To fully understand the role of scope in the tractor's diffusion, it is necessary to first understand how tractors are used and the associated technical demands. Tractors power and tow the agricultural implements that do the day-to-day work of plowing, planting, cultivating, and harvesting crops. Given the diverse demands for power in farming, the modern tractor is no small feat: the technical requirements vary not only across stages of crop production, but also across different crops, especially for crops grown in dense fields versus organized rows. Early models could not navigate row crops for cultivation and harvest without destroying the crop, and this generation of tractor technology was therefore not a candidate to replace draft power on corn farms at *a*ny price. By the 1930s, however, more versatile designs emerged and made it possible for these farms to "replace [all] their horses and mules with one general-purpose tractor" (Sanders 2009).

As a direct consequence of this path for product development, early tractor diffusion was overwhelmingly concentrated in the Wheat Belt states of North Dakota, South Dakota, and Kansas. As contemporaries noted, "the possible market for tractors ... in the corn belt has hardly been scratched" (Iverson 1922). Between 1930 and 1940, this pattern reversed, with diffusion proceeding most quickly in Corn Belt. This historical sequence is plainly visible in maps of wheat acreage, corn acreage, and diffusion in 1930 and from 1930-40, shown in Figure III.1.

Figure III.1: Tractor diffusion in U.S. Midwest, 1930 and 1930-40




Notes: Figure shows the distribution of wheat and corn production in the U.S. Midwest (left) and tractor diffusion through 1930 and from 1930-40 (right). Crop concentrations calculated as the fraction of farmland in the given crop; tractor diffusion as the fraction of farms owning a tractor. Darker blues represent higher values. Counties in white omitted due to missing data or because their borders changed over the sample period. Data from 1910-1940 Census of Agriculture.

Traditional explanations for technology diffusion cannot fully account for these patterns. Wheat Belt farmers were not less credit-constrained in the 1920s – if anything, these areas were in greater financial distress following the post-World War I collapse in wheat prices. Though in 1930 there was a cluster of diffusion in eastern Illinois, nearer to the major manufacturers of the day (in Chicago, Milwaukee, and Detroit), the majority of diffusion occurred in the region's western periphery, suggesting only a limited role for trade costs and information spillovers. And while farms were substantially larger across the Plains relative to other parts of the Midwest, fixed costs and acreage thresholds cannot explain the gap that persisted in Nebraska. The source of these patterns ultimately lies in the path of manufacturers' R&D in tractor technology.

Using county-level data on farm characteristics, and exploiting the fact that crop choices persistently reflect exogenous climatic and soil conditions, I estimate the relationship between tractor diffusion in 1930, and the change in diffusion from 1930-1940, and county crop mix. The results confirm that tractors were significantly quicker to diffuse to wheat-growing regions of the Midwest relative to corn-growing regions, and that this pattern is not the result of New Deal relief, the Dust Bowl, financial conditions, local factor prices, or other variables that could conceivably determine tractor demand. The estimates suggest that relative to state averages, tractor diffusion in 1930 was on average 5 percentage points (p.p.) greater for every 10 percent of farmland planted in wheat but did not vary with acreage planted in corn. By 1940, diffusion was still roughly 5 p.p. higher for every 10 p.p. in wheat, but it was also nearly 4 p.p. greater for every 10 percent of farmland in corn, having caught up over the intervening decade. Under counterfactual projections, aggregate diffusion throughout the Midwest would have been 25 percent higher in 1930, and approximately equal across all of the major grain-growing states in the region (and country), had tractors diffused at the same rate to counties with equal concentrations of wheat and corn.

Given the tractor's historical impact on agriculture, an increase or temporal shift of this magnitude would have had significant consequences for agricultural productivity and output. The tractor completely upended the organization of U.S. agriculture, reducing labor requirements by at least one-fourth through 1960, and increasing available cropland by even more (Olmstead and Rhode 2001). What was once slow, back-breaking work for humans and animals alike became relatively effortless. As Olmstead and Rhode (1994) put it, "the conversion from draft power to the internal combustion engine was one of the most far-reaching technological changes ever to occur in the United States." Confirming this sentiment, Steckel and White's (2012) estimates suggest that by 1954 the tractor was generating social savings of as much as 8.6 percent of GNP.

Though the tractor is important in its own right, I argue that scope is an important, inherent property of technology diffusion. To support this claim, I re-examine the diffusion of steam and electric power in U.S. manufacturing and hybrid corn in agriculture. Each of these technologies first developed for narrow applications – the steam engine for pumping water out of coal mines; electric motors for use in electric street cars, subway trains, and elevators; and the earliest commercial varieties of hybrid corn for the heart of the Corn Belt – and only after the technology generalized did it become truly pervasive. Moreover, lags in scope could be as great or greater than lags in scale: the time it took to go from a positive level of hybrid corn diffusion in one state to positive levels in all 48 contiguous states exceeded the time it took to go from 0 to 100 percent diffusion within several of these states individually. Using state-level data on acreage planted to hybrid seed, I show that diffusion follows an S-shaped pattern not only along the well-documented intensive margin, but also along its extensive margin – confirming the model's predictions.

A natural implication of these results is that in addition to studying the population of users, researchers and policymakers should also focus attention on the firms performing R&D that increases the scope of existing technologies such that they can be used more broadly. Given the presence of externalities that decouple private returns to R&D from social returns, a second implication is that investment in technological generality may be a high-value target for R&D policy tools. The results of this paper might also be able to explain previously-documented spatial patterns in technology diffusion, such as the evidence from Comin and Hobijn (2004) that technology diffusion "trickles down" from more- to less-developed economies and from Keller (2002) that R&D spillovers appear to decline with distance: new technology is often first developed in more advanced regions and in many cases would have to be adapted to conform to local conditions, users' needs, and technology standards in other parts of the world in order to penetrate these markets.

The paper proceeds as follows. In Section III.1, I present a model of R&D in specific- and general-purpose technological attributes, obtain predictions for the path of product development and scope of diffusion, and derive policy implications. In Section III.2, I review the tractor's history from the 1890s to the 1940s, drawing on narrative accounts, with an emphasis on its expanding capabilities. In Section III.3, I describe the data and empirical strategy for identifying the scope of tractor diffusion in 1930 and 1940. In Section III.4, I estimate the relationship between crop mix and tractor diffusion in the U.S. Midwest, provide counterfactual projections, and briefly discuss diffusion in other regions. In Section III.5, I show that scope is equally important to explaining lags in the diffusion of steam engines, electric power, and hybrid seed corn. Section III.6 concludes.

## III.1 Theoretical Framework

Suppose a monopolist inventing firm develops a technology that it sells to users in an arbitrary number of application sectors. The focal technology is characterized by general-purpose quality $z_g$ and a vector of application-specific qualities $\{z_a\}_{a\in A}$ across a range of applications $a \in A$, with associated R&D costs $C^g(z_g)$ and $\{C^a(z_a)\}$, which are increasing and convex in their arguments. General-purpose quality is embodied in features that are generic or useful for many purposes, such as the rotary motion produced by a motor. Application-specific quality is embodied in features which are limited in scope, like a component that performs a specific, repetitive task, and as the terminology implies is valuable only in particular applications of the technology. Within such applications, this limited functionality can substitute for general functions (e.g., a self-powered component). The technology's total quality in a given application $a$ can thus be expressed as $\zeta_a(z_g, z_a) = z_g + z_a$. In this framework, general quality is special for two reasons: it is useful across many applications, and it complements the sector-specific technologies of other firms. I assume that the focal technology is produced at marginal cost $c$ and sold at price $w$.

Developers in the application sectors create complementary products that address a narrow, sector-specific need. Each such product is characterized by quality $T_a$ and increasing and convex R&D costs $C^a(T_a)$. Application sector firms' investment in quality improvements generates private returns of $\Pi^a(T_a|w, z_g, z_a)$. The exact form of $\Pi^a$ depends on the downstream market structure and is nonessential; the key assumptions for the purposes of this paper are (1) that $\Pi^a$ is decreasing in $w$ and increasing in $z_g$, $z_a$, and $T_a$, and (2) that $\Pi^a_{z_g T_a} \geq 0$. This latter assumption implies innovational complementarity between the focal technology and sector-specific complements, as in Bresnahan and Trajtenberg (1995): improvements in the former make complementary innovation more profitable, and vice versa. Changes in the application-specific quality of the focal technology for other applications (i.e., changes in $z_{\tilde{a}}$ for $\tilde{a} \neq a$) have no direct bearing on $\Pi^a$.

It may be useful to elaborate on these ideas with a concrete example. The modern farm tractor is a device with substantial general-purpose quality: its primary functions are to power and tow the agricultural implements used for field operations, and it is now used with nearly all crops and in nearly all stages of crop production, from pre-planting to harvest. Indeed, the earliest tractors with this versatility were aptly marketed as "general-purpose" tractors. Tractor-drawn implements, on the other hand, are used for a single task or stage of crop production and may even be specific to a single crop. Implements are sector-specific technologies that complement, and are complemented by, the tractor's various features. Tractor-drawn implements can be contrasted with independent devices such as the standalone cultivator or grain combine, each of which is used in a single stage of production and neither of which requires a separate tractor. In this model, these could be described as tractor-like devices with a high level of application-specific quality.

Each application sector is assumed to include a single sector-specific inventing firm. These firms undertake R&D to maximize firm-specific profits $\Pi^a$, subject to a periodic budget constraint $B_a$. Within this framework, firm $a$'s solution is to expend its R&D budget each period developing $T_a$ up to the point where the marginal returns to R&D equal the incremental cost. Denote this solution as $T_a^* = T_a^*(w, z_g, z_a)$. This function is increasing in $z_g$ and $z_a$, which can complement $T_a$, and decreasing in $w$, which reduces demand for the focal technology and in turn its complements. The presence of a budget constraint does not change this solution, but it does introduce the possibility for delays: difficult or expensive R&D will slow down product development. Although this feature isn't crucial to what follows, it is useful for explaining historical episodes in which the development of complementary equipment lags behind that of the focal technology.

Demand for the focal technology from each sector $a$ takes the form $X^a(w, z_g, z_a, T_a^*)$, where $X_w^a < 0$, $X_z^a > 0$, $X_T^a > 0$, and $X_{wz}^a < 0$ for $z \in \{z_g, z_a\}$. It follows that the focal firm's marginal revenue has similar properties. The latter condition implies that the firm supplying the focal technology "cannot appropriate more than the incremental surplus" generated by quality improvements (Bresnahan and Trajtenberg 1995), leading it to undersupply quality. In essence, whenever the firm invests in quality improvements, a fraction of the ensuing rents will accrue to inventors of complements, and these rents cannot be fully re-appropriated: if the firm attempts to tax these developers (e.g., with licensing fees) to re-appropriate this surplus, it will reduce their incentive to invest in $T_a$, and demand will accordingly decline. As a result, the focal firm's investment in expanding the scope of its technology will be less than the social optimum.

## The Path of Product Development

The focal inventing firm must choose how much general-purpose and application-specific quality to develop each period, subject to its own periodic R&D budget constraint $B_g$. If the returns to application-specific quality or the costs of developing general-purpose features are large, or if complementary technologies exist for only a handful of applications, the firm may prefer to invest in specific features in advance of more general features. Formally, the firm's problem is

$$\max_{z_g, z_{a_1}, \ldots, z_{a_n}} \Pi^g(z_g, z_{a_1}, \ldots, z_{a_n} | c, \mathbf{T}) - C^g(z_g) - \sum_a C^a(z_a) \ ,$$

where

$$\Pi^g(z_g, z_{a_1}, \ldots, z_{a_n}|c, \mathbf{T}) = (w^* - c) \sum_{a \in A} X^a(w^*, z_g, z_{a_1}, \ldots, z_{a_n}, T_a)$$

is the firm's return, $w^*$ is the monopoly price, and $\mathbf{T}$ is a vector of complementary technologies' quality. Note that since the firm takes price $w^*$ as given when solving for the profit-maximizing levels of $z_g$ and $\{z_a\}$, the assumption of monopoly is unnecessary as long as the firm can (even temporarily) retain rents from innovation. The firm's solution is $\mathbf{z}^* = \{z_g^*(c, \mathbf{T}), z_{a_1}^*(c, \mathbf{T}), \ldots, z_{a_n}^*(c, \mathbf{T})\}$. Due to innovational complementarities, each $z^*$ is increasing in $\mathbf{T}$. As with the application sector firms, the budget constraint dictates the pace at which the firm converges to this solution – but not the form this solution takes – and explains why product development is not instantaneous.

The long run solution has $\Pi_\gamma^g / C_\gamma^\gamma = 1$ for all $\gamma = g, a_1, \ldots, a_n$. But as long as the R&D budget constraint is binding – in other words, at all points along the adjustment path – the features with the highest shadow price will be developed until others exceed them. In practice, this means that the focal technology will often first develop for applications with exogenously high demand or exogenously inexpensive development costs. History suggests there is merit to this argument, as many technologies that eventually evolved to be very general first developed with applications limited to areas in which they were especially needed – including the primary examples (tractors, steam engines, electric motors, hybrid corn) discussed in this paper. Only when the gains to specialization are exhausted will product development proceed to general-purpose features, if at all. A typical path for product development will therefore be:

1. Invention for applications with exogenously high demand or inexpensive R&D
2. (Potentially) Develop general-purpose capabilities that serve a wide range of users
3. (Potentially) Round out development of remaining application-specific features

These results can be summarized with the following proposition.

**Proposition 1.**

*1) In the long run, general and application-specific quality will develop up to an interior solution where marginal benefits equal costs across all $\gamma = g, a_1, \ldots, a_n$ for which $z_\gamma > 0$.*

*2) Product development will follow an adjustment path along which technological attributes with the highest shadow price are developed until others exceed it; in practice, this means that the focal technology may first develop for particular applications and only later generalize to broader use.*

## Implications for the Scope of Diffusion

The predictions for the path of development are intuitive: product features develop in the order in which they are most valuable. In some cases, product development will lead to a general-purpose variant, but in many cases it may never get there. Externalities can also constrain investment in greater scope by driving a wedge between private and social returns to innovation, implying that there is a role in this setting for well-designed R&D policies. But the most striking result from the model is its implication for diffusion, and in particular for understanding the source of cross-sectoral lags in diffusion, which will be shaped by the set of applications for which a given technology can be used at all. Since scope must precede scale, this margin can play a paramount role in explaining diffusion lags both in cross-section as well as in the aggregate.

Griliches (1957) calls it the "availability" problem: the diffusion of hybrid corn within a given state or crop reporting district required a seed variety adapted to local growing conditions. I refer to it as the scope, or extensive margin, of diffusion. The key insight is that cross-sectional variation in diffusion results not only from the rate at which it proceeds, but also from when it begins. And because product development naturally proceeds from specific- to general-purpose variants, diffusion will tend to follow a characteristic S-curve not only within applications, but also across them. Later sections verify this prediction for hybrid corn: for any fixed level of diffusion, and in particular for lower levels indicating availability of locally-adapted varieties, the number of states that have surpassed that level of diffusion follows an S-shape over time.

The argument can be formalized as follows. Recall that the focal technology has quality $z_g + z_a$ for applications in sector $a$. We can write diffusion in sector $a$ as $D_a = F(z_g + z_a)$, where $F(\cdot)$ is a characteristic S-shaped CDF. Diffusion in sector $a$ is thus increasing in both $z_g$ and $z_a$, while diffusion in sector $b$ will be increasing in $z_g$ but not directly affected by $z_a$. Since $F(\cdot)$ is one-to-one, it has a functional inverse $F^{-1}(\cdot)$, and we can write

$$\begin{bmatrix} F^{-1}(D_{a_1}) \\ \vdots \\ F^{-1}(D_{a_n}) \end{bmatrix} = \begin{bmatrix} z_g + z_{a_1} \\ \vdots \\ z_g + z_{a_n} \end{bmatrix} . \tag{III-1}$$

Equation (III-1) is a system of $n$ equations with $(n+1)$ unknowns, one of which may be normalized to zero with no loss of generality. As $z_g$ increases, diffusion will increase across all applications sectors, including previously untapped markets, and as product development evolves from specific- to general-purpose features, so will the scope of diffusion begin with a narrow set of applications, accelerate to many others, and then top off with the remainder.

**Proposition 2.** *The scope of diffusion varies one-to-one with that of R&D. Diffusion will therefore typically follow an S-shaped pattern not only within applications, but also across them.*

In concept, the diffusion of the focal technology should also depend on the quality of complementary innovation. This parameter is omitted from equation (III-1), since it is fully determined by the characteristics of the focal technology itself (recall that $T_a^* = T_a^*(w, z_g, z_a)$). What this implies in practice is that when the focal technology improves in its general-purpose capabilities, complements should immediately develop to take advantage of these new features. Historical experience broadly concords, though exceptions do exist. For example, the mechanical corn harvester was invented just five years after the the first general-purpose tractor (1930 and 1925, respectively). Firms similarly began attacking the cotton harvesting problem immediately following the development of a general-purpose tractor, but the mechanical cotton picker was relatively slower to develop due to the difficulty of the engineering problem as well as institutional features of the U.S. South constraining demand (Whatley 1985, 1987).

## III.2 History of the Tractor

The modern tractor's history begins around 1870 with the invention of the steam tractor, which was effectively little more than a steam engine on wheels. These were equipped with a drawbar for towing portable implements and a belt pulley to power stationary equipment, and were primarily used for plowing and post-harvest threshing, with little portable use beyond tillage. They were also heavy, expensive to purchase and maintain, and prone to mechanical failures and explosion. While they were never a serious threat to farms' dependence on draft power, steam tractors were a clear antecedent to the internal combustion tractors pervasive in agricultural today.

Kerosene tractors were developed around 1890 yet were hardly an improvement on steam models. As Olmstead and Rhode (2001) put it, the earliest gas tractors were expensive behemoths, much like the steam tractors that preceded them, and had similarly limited functionality; any portable implement that needed to be powered would either have to get that power from the movement of a bullwheel or provide it independently. Given their immense size, cost, and unreliability, tractor diffusion was practically nonexistent prior to 1910 (Figure III.2). The transition to small, lightweight, affordable tractors began with the Bull tractor in 1913, but this transition was only finalized with the commercial introduction of the Ford Fordson four years later, in 1917.

[Figure III.2 about here]

The Fordson was the first big commercial success in the tractor industry, and by all accounts – including Figure III.2 above – it marked the beginning of the tractor era. By the end of 1918, Ford had overtaken its competitors in sales (Leffingwell 1998), and by the early 1920s, the Fordson accounted for 75 percent of all tractor sales in the U.S. (Leffingwell 2002). Henry Ford continued his assault on the tractor industry by initiating a price war in 1922, cutting the Fordon's price by 35 percent overnight. The "tractor price wars" (as they are now called) led to a wave of consolidation from over 150 manufacturers to just a few dozen. By the time Ford ended production of the Fordson in 1928, it had sold nearly half of *a*ll tractors sold in the 1920s (White 2010).[4]

The advantage of the Fordson was its size, agility, and low price, but its low clearance made it impractical for cultivating row crops such as corn or cotton, leading manufacturers to separately develop and sell expensive, standalone cultivators (Sanders 2009) and Corn Belt farms to continue relying heavily on draft power. Contemporary observers noted that "The possible market for tractors ... in the corn belt has hardly been scratched, for study reveals that only about six per cent of the farms in these six states have tractors, while the other ninety-four percent still depend on horses for power" (Iverson 1922). The "logical solution" was to "design a tractor that will do cultivating as well as plowing, disking, dragging, and other drawbar work."

IHC saw these deficiencies as an opportunity to develop a "general-purpose" (G-P) model, and in 1925 it began selling the first such tractor – the Farmall.[5] The Farmall had high clearance and adjustable-width treads for use in all of plowing, cultivating, and harvesting, on both row crops and

[4] Other sources agree that Ford dominated the decade: Gilbert's (1930) survey of four agricultural regions in New York in 1926 revealed that 54.7 percent of tractors used on surveyed farms were Fordsons.

[5] The IHC effort to develop a general-purpose tractor was spearheaded by a single employee, Bert Benjamin, and supported by the engineering department. According to Benjamin, while "there was talk about a new kind of tractor in the industry" at the end of the 1910s, it was also the case that "no one had such a machine or even much of an idea on how to start building one" (Klancher 2008). The first references to this project in IHC records appears in

small grains. It also had a more powerful engine, a belt pulley to power stationary equipment, and a motor-driven shaft that could power implements (power take-off). As Sanders (2009) describes it, "It was designed (and thus named) to accomplish all of the power needs on the farm. At last, farmers could replace their horses and mules with one general-purpose tractor."

The Farmall was an instant hit, and it ushered in a new generation of tractor technology as competitors rushed to imitate the Farmall's design and develop their own G-P tractors. John Deere came out with a version in 1928, and Allis-Chalmers in 1930, but by that point the Farmall was already dominant, having overtaken Fordson sales in 1927/28. Further advances in tractors soon followed: in 1927, Deere invented the power lift for raising implements during turns – an enervating and time-consuming task; in 1931, Caterpillar built the first diesel-engine tractor; in 1932, Allis-Chalmers introduced pneumatic rubber tires that improved fuel efficiency and forward horsepower; and in 1938, Ford introduced the Ferguson three-point hitch for attaching implements, replacing the drawbar. Manufacturers quickly made these features standard, and by the early 1940s the industry had arrived at a dominant design: the main features of the modern tractor had been set. Over the following decades, G-P tractors "would change little, except for increasing in size and horsepower" (White 2010) and adding comfort and safety features.

### History in Relation to the Theory

Several features of the model in Section III.1 are embedded in this history, especially the sequence of generalization and the co-development of tractors and complementary agricultural implements. The tractor's earliest applications were in tillage – back-breaking work for animals and humans alike. The physical requirements of plowing generated exogenously high demand for mechanical power and explains why the first steam tractors were developed to be, and termed, "plowing engines." Tractor development subsequently continued in applications related to grain production, where demand was relatively high, the engineering problem was relatively easier, and complementary harvesting equipment was already available. Only later, when the marginal gains to improvements in standard tractors were exhausted, and specialized equipment such as standalone cultivators were found to be unprofitable, did manufacturers direct their research effort towards a general-purpose design with broad-based demand – and its diffusion rapidly followed.

Once the tractor generalized, implements were invented to perform nearly any task in the field. Plows, harrows, planters, and grain harvesters, threshers, and combines were all available for use with standard, Fordson-type tractors. Later came cultivators, corn harvesters (1930), cotton pickers

---

1919, but it took many experimental prototypes, each built at considerable cost, to arrive at a commercially viable product. The earliest prototypes looked little like other tractors of the day and were heavy and cumbersome to operate, but they could be used with many more implements than the firm's existing models. Company records from 1921 reveal that executives were unenthusiastic about these early prototypes and many wanted to abandon the research program; it continued in large part due to the support of the firm's president. By the end of the year, Benjamin had proposed a version that looked more like a Fordson but had higher clearance and a narrow front wheelset – beginning to resemble the model that would eventually be commercially sold – and support for the idea within the company was beginning to grow. By 1923, the engineering group had shifted to making more marginal improvements in power, weight, and cost, and in 1924, the firm began a production run of 200 units, which it quietly sold without advertising in case they proved unsuccessful and to avoid cannibalizing existing product lines. Further tweaks were implemented in 1925, as sales gradually expanded, and in 1928, when Ford exited the industry, the Farmall was there to take its place. For a more detailed history, see Klancher (2008).

(1942), and harvesters for several other crops.[6] Mechanical corn harvesters entered production only a few years after the Farmall, supporting the theoretical assertion that manufacturers of complementary devices respond quickly to improvements in general-purpose functionality. IHC similarly began working on a mechanical cotton picker immediately after the Farmall and by the early 1930s believed it had solved the fundamental engineering problems for such a device (Whatley 1987, referencing the IHC "New Works Committee"). Tractors were in turn adapted to be used with such equipment, reflecting the two-way effects of innovational complementarities, and improvements such as power take-off and the three-point hitch specifically served this end.

Most importantly, although tractor diffusion began in the late 1910s, it was apparent even to contemporaries that this early diffusion was restricted in scope by the tractor's limitations, which were only overcome by general-purpose models. Figure III.2 reveals that the data accord with this narrative history: 1920-30 is the first decade of the tractor, with diffusion rising from 3.6 to 13.5 percent of U.S. farms, mostly in Midwest and Pacific states. Figure III.3 below shows the path of diffusion in each Midwest state from 1920 to 1940. In 1920, only 6.8 percent of farms in the Midwest owned a tractor. Between 1920 and 1930, this fraction nearly quadrupled to 25.7 percent of Midwestern farms, and by 1940, diffusion had reached 42.4 percent, despite stagnating during the Great Depression. North Dakota, South Dakota, and Kansas led the trend towards mechanization throughout the 1920s, while Corn Belt states like Iowa, Illinois, and Nebraska were relative laggards in 1925/1930 but experienced the largest increases the following decade.

[Figure III.3 about here]

### Previous Research on Agricultural Technologies

Though a large body of research has examined the historical diffusion of tractors and other agricultural technologies, the distinction between scale and scope is missing from this literature. Most research treats the tractor as a product of uniform quality both over time and across space and attributes lags in diffusion to fixed costs with indivisibility, credit constraints, or exogenous factor price changes. Even when the existing literature recognizes that "a 'tractor' in 1960 is not the same capital good as a 'tractor' in 1920" (Manuelli and Seshadri 2014), it tends to overlook the fact that tractor quality varied as much or even more in cross-section as it did over time.

David's (1966) study of antebellum reaper adoption introduced the neoclassical threshold model to this literature, asserting that reaper diffusion was driven by increases in farm size. Olmstead (1975), however, calls into question the assumption of a static, indivisible technology, showing that joint ownership and contract work were common practice and that reaper quality was improving, and suggests that farm size was in fact simultaneous with the adoption decision. Ankli and Olmstead (1981), Clarke (1991), White (2000), and others have nevertheless attempted to calculate adoption thresholds for tractors in order to explain its delayed diffusion, despite the well-known critiques of David's (1966) model. Myers (1921) and Gilbert (1930) lend support to both advocates and critics of the threshold model, acknowledging that "the advantages of a tractor increase with [the] size of the farm" while also pointing out that contract work was common and that tractor adoption

[6]Harrows smooth the soil after plowing, planters lay the seed, and cultivators turn the topsoil for secondary tillage of budding row crops. Grain harvesters, threshers, and combines cut grain stalks and separate the grain itself from the chaff. Other mechanical harvesters are generally crop-specific. Mowers and balers cut and bundle hay.

led farms to expand: "the ability to do more work with the tractor resulted in an increase in the amount of land worked on nearly one-third of the farms visited" (Gilbert 1930).

Clarke (1991) argues that financial barriers slowed tractor diffusion in Illinois and Iowa in the 1920s and that New Deal relief – rather than changes in farm size, factor prices, or technology – was responsible for a surge in diffusion in the 1930s. To support this claim, Clarke first calculates a 1929 adoption threshold of 100 acres for farms in Corn Belt states. Clarke then finds that only about half of the farms above this threshold owned a tractor in 1929, and that this gap narrowed over the subsequent decade. After correlating "underdiffusion"[7] with farmers' cash holdings and mortgage debt ratios and obtaining coefficients with the expected signs (negative and positive, respectively), she attributes the growth in diffusion to New Deal price supports and lending programs that might have improved Corn Belt farmers' financial positions and borrowing conditions.

Would-be adopters would have had to be credit-constrained for New Deal policies to cause a surge in tractor purchases. Yet farms in North and South Dakota were leading adopters of tractors in the 1920s, despite the post-WWI collapse in wheat prices and mortgage foreclosure rates near 50 percent (Alston 1983, Table 1). White (2000) further notes that "the same farmers that Clarke concluded might not have been able to obtain a loan for a tractor were cheerfully buying automobiles for cash" before 1930: roughly 80 percent of farms in Midwest states owned automobiles at that time, compared to only 25 percent with tractors. The difference was not for a lack of manufacturer credit, as both Ford and IHC provided financing to their customers. Given these inconsistencies, the evidence that liquidity constraints can explain diffusion lags in the Corn Belt is questionable, though financing undoubtedly plays an important role in large equipment purchases.[8] In Appendix III.F, I use data on New Deal relief at the county level (borrowed from Fishback, Kantor, and Wallis 2003) to explore the possibility that improving financial conditions were more important to 1930s tractor diffusion than the technical advances that are the focus of this paper.

Manuelli and Seshadri (2014) counter the claim that tractor diffusion was inefficiently slow due to market imperfections such as credit constraints with the more traditional argument that exogenous changes in factor prices and improvements in tractor quality over time can rationalize the tractor's allegedly slow diffusion. Accounting for the tractor's changing quality over time is an important addition, but by modeling only aggregate diffusion and ignoring the variation in quality across space, it misses a crucial part of the story: tractors hardly diffused to farms growing row crops until the 1930s because they could not replace the horse *at* any price. Treating the tractor's quality as a unidimensional parameter that follows a secular process over time, and using it to explain the scale of diffusion at the aggregate level, belies the true nature of the problem.

[7]Defined in Clarke (1991) as the fraction of farms above the 100-acre threshold without tractors.

[8]Clarke's regressions also suffer from ordinary econometric shortcomings, as adoption thresholds are simultaneous with financial conditions: an increase in the interest rate will necessarily raise the threshold, resulting in a mechanical decline in diffusion among 100+ acre farms, as farms just above 100 acres will no longer be in the adoption zone. Large and small farms may also be differentially likely to be mortgaged – an additional source of simultaneity.

## III.3 Data and Empirical Strategy

For empirical evidence, I turn to a panel of 1,059 counties in the U.S. Midwest through 1940.[9] The Midwest led the country in tractor adoption through WWII and exhibits sufficient spatial variation in diffusion early on in the tractor era to discern its expanding scope. The Midwest also spans the principal grain-producing U.S. counties, making it of inherent interest.

This analysis integrates data from several sources. I use county-level data in Midwest states from the 1910, 1920, 1930, and 1940 U.S. Census of Agriculture to measure tractor diffusion, investment in agricultural implements, farmland, crop mix, and other characteristics of farms and farmers. I draw on the U.S. Census of Population in the same years for supplementary county-level data. The dataset also includes records of bank failures from the FDIC; county-level New Deal expenditures from Fishback, Kantor, and Wallis (2003); point-to-point freight rates and local railroad density from W.J. Hartman Publishing (1916) and Donaldson and Hornbeck (2013), respectively; Dust Bowl soil erosion from Hornbeck (2012); average levels and variation in elevation and rainfall at the county-level, from the USGS and PRISM Climate Group at Oregon State University, respectively; and usual harvesting dates by state and crop, from the USDA. I use these data to understand and control for other features of U.S. agriculture affecting tractor adoption.

I use the NHGIS county boundary shapefiles (Minnesota Population Center 2011) for the 1910-1940 Census years to aggregate continuous geospatial data (elevation, rainfall) at the county level and drop all counties that merged or divided over the sample period as well as counties whose geographic centroids shifted more than 0.01 degrees in latitude or longitude between decades. The main analysis treats remaining counties' borders as static, reflecting the stability over these years of the centroids calculated by mapping software. For robustness I explore the sensitivity of the analysis to adjustments that maintain 1910 borders, following the procedure described in Hornbeck (2010), and obtain statistically and quantitatively similar results.

### Empirical Methods and Identification

Identification hinges on the fact that particular areas are inherently better suited to growing different crops for exogenous reasons (such as soil type, climate, etc.), and an assumption that farmers' crop choices reflect these local advantages regardless of whether they own a tractor. If the historical account is true, diffusion in the 1920s should be higher in areas where field grains are grown and lower in areas more concentrated in corn. Following the development of the general-purpose tractor in the late 1920s, the difference should then mitigate or reverse, with corn-heavy counties experiencing catch-up increases in diffusion. Formally, the identifying assumption is that crop mix is independent of any unobserved factors affecting the decision to adopt the tractor. Since crop mix is simultaneous with the choice of inputs and technology, I also instrument for contemporaneous

[9] The included states are: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin. These states form the East and West North Central Census Divisions. Across these 12 states, there were 1,049 counties in 1910, 1,056 in 1920, 1,057 in 1930 and 1940, and 1,056 in 1950, reflecting county mergers and divisions over time, mostly occurring in North and South Dakota. In sum there are 1,059 uniquely defined counties over this period. The baseline sample is restricted to the 1,035 counties whose borders were unchanged from 1910 to 1940; regression sample sizes are generally less than 1,035 due to missing variables for some county-years. I forgo Hornbeck's (2010) border adjustment procedure in the main analysis to avoid synthetic observations constructed by piecing together fractions of counties on the assumption that county-level variables are evenly dispersed across space, though the results are insensitive to this choice.

crop mix with pre-tractor values and regress directly on lagged values, as tractors were unlikely to have influenced crop choices in 1910 (or even 1920, in some areas), when there were only around 1,000 tractors on U.S. farms (Historical Statistics, Table Da623).

Table III.1 provides more information on the mix of crops grown in the Midwest through World War II. Six crops – corn, wheat, oats, barley, rye, and hay – alone accounted for roughly 50 percent of farmland and 85 percent of harvested acreage in the Midwest. Though wheat acreage temporarily spiked at the end of the first World War, and acreage in oats and hay declined modestly along with the draft stock it fed, the fraction of Midwest farmland planted in each of these crops is on the whole fairly steady from 1910 to 1940. Appendix III.B provides further evidence that their level and distribution as a fraction of farmland, cropland, and harvested acreage changed little over the period. The most compelling evidence of a stable Wheat Belt and Corn Belt can be seen in the maps in Appendix III.G, which plot the spatial distributions of acreage in each crop and confirm that these distributions are unchanged throughout the sample period.[10]

[Table III.1 about here]

## III.4 The Scope of Tractor Diffusion

According to the historical narrative, "the tractors of the Fordson generation were ill-adapted to the majority of farms in the United States [and] were of questionable value on farms growing corn or cotton or other row crops" (Williams 1987). As such, tractor diffusion in the 1920s was by and large confined to farms growing small grains: until the general-purpose tractor was invented, it was necessary to keep horses for cultivating and harvesting corn, and the cost of owning both a tractor and a team of draft animals was prohibitive to nearly any farm. This can be seen most directly in Figure III.1 (in the introduction), which shows the geographic distribution of wheat acreage in 1910 and tractor diffusion in 1930. Counties where tractors were most common in 1930 tended to be counties where wheat was historically grown in higher concentration.[11]

The development of the general-purpose tractor and other technological advances rapidly transformed the industry: Fordson-type tractors' share of sales for domestic use declined from 92 percent in 1925 to 4 percent in 1940, while G-P tractors' share grew to 38 percent in 1935 and 85 percent by 1940 (Olmstead and Rhode 2001). These technical advances allowed the tractor to completely replace draft animals in corn production, and the tractor's subsequent diffusion to the Corn Belt between 1930 and 1940 is as visible in Figure III.1 as its absence pre-1930.

The visual evidence in maps is corroborated by descriptive statistics. Table III.2 compares tractor diffusion in counties above and below the median percent of farmland in corn, wheat, and barley in

[10]The Midwest crop mix began to change in the early 1940s, shortly after the sample ends. Prior to the 1940s, soybeans were uncommon and were typically grazed off or plowed under after being used for nitrogen fixation. Harvested soybean acreage exploded during the war, much of it displacing corn fields in the Corn Belt. The other new crop with significant acreage was sorghum, mostly grown in Kansas, Oklahoma, and Texas. The current acreage and geographic concentration of field crops is available from USDA (2010).

[11]A cluster of diffusion is also visible within a roughly 100-mile radius of Chicago. Diffusion in these counties may have been influenced by their proximity to IHC, which had its headquarters in and shipped from Chicago. Proximity to Chicago would have reduced the freight costs paid by purchasing farmers, which could range from a few percent to upwards of 15 or 20 percent of the sticker price in the outer reaches of the Midwest (calculations based on data from Hartman's 1916, Donaldson and Hornbeck 2013). IHC also had a strong, century-old marketing presence in northern Illinois, which may partially explain why farmers in these counties adopted early.

the sample. By 1930, farms in counties with an above-median concentration of wheat had adopted the tractor at a significantly faster rate than those in below-median counties, and an even stronger pattern holds for barley. These patterns persist through 1940 and are significant at the one percent level. In stark contrast, farms in counties with high and low concentrations of corn were using tractors at statistically identical rates in 1930, and only over the next decade did diffusion in above- and below-median counties begin to diverge. By 1940, tractor diffusion was quantitatively the same in counties with a high concentration of either wheat or corn.

[Table III.2 about here]

To empirically identify the tractor's evolution into an all-purpose machine, and the consequently expanding scope of its diffusion, I regress county-level tractor diffusion in 1930 and 1940 and the change in diffusion between 1930 and 1940 on the local crop mix and controls. I present two sets of regressions: Table III.3 shows the results from regressions of diffusion on contemporaneous crop mix, and Table III.4 instruments with 1910 values, which pre-date the tractor era and reflect the naturally-occurring comparative advantage to growing wheat (or corn, etc.) in the Wheat and Corn Belts.[12] The principal estimating equation throughout this section is the following:

$$\text{Diffusion}_{it} = \beta_0 + \sum_{c=1}^{5} \beta_c \text{CropPctOfFarmland}_{c,it} + \mathbf{X}_{it}\theta + \gamma_s + \varepsilon_{it}$$

where $i$ indexes counties, $t$ indexes years, and $c$ indexes crops. The crop mix variables are calculated as the fraction of county farmland in each of corn, wheat, oats, barley, and rye. The $\mathbf{X}_{it}$ term represents a set of county-level controls, and $\gamma_s$ denotes state fixed effects.

In each table I provide six specifications. All specifications include state fixed effects, and the latter three columns of each table additionally control for county longitude and latitude, average rainfall, elevation, and distance from the f.o.b. shipping terminals of Ford (Detroit) and IHC (Chicago). These specifications also control for farm size, local interest rates, farm mortgage debt ratios, horses per acre of farmland (contemporaneous and 1910), mules per acre (contemporaneous and 1910), and per-acre expenditure on labor in 1910 from the Census of Agriculture.

The OLS estimates in Table III.3 shows that farms in counties growing more wheat, oats, and barley were significantly more likely to have adopted tractors in 1930, while those in counties concentrated in corn or rye adopted tractors with much lower frequency. However, by 1940, counties with similar concentrations of wheat and corn were using tractors at similar rates.

[Table III.3 about here]

These patterns are confirmed in the IV regressions of Table III.4, where the coefficients for wheat and corn are quantitatively similar to the OLS estimates: wheat-growing counties adopted tractors at high rates before 1930, and diffusion in corn-growing counties caught up over the subsequent decade. The table also provides the minimum F-statistic from the first-stage regressions for the crop mix, confirming that crop mix in 1910 strongly predicts that in 1930 or 1940.

[12] In unreported tables, I regress directly on these lagged values and find similar results.

[Tables III.4 about here]

The 1940 Census of Agriculture provides detailed information on farms' ownership of tractors by model vintage, reporting the number of farms whose latest model-year tractor is pre-1930, 1931-1935, and 1936-1940. Tables III.5 and III.6 regress the share of each vintage among farms reporting any tractor vintage in 1940 on the crop mix and control variables used in the previous regressions.[13] As with the diffusion regressions, I provide two variants (OLS and IV). OLS estimates reveal that counties concentrated in corn, oats, barley, and rye were significantly more likely to own later-vintage tractors in 1940 relative to the average county, while the opposite is the case for wheat. The IV estimates provide similar results but emphasize that the tendency to own the latest-vintage tractors was limited to counties with high concentrations of corn or barley.

[Tables III.5 and III.6 about here]

In addition to lagging diffusion in the Corn Belt, two others results stand out across all tables. First, barley-heavy counties both adopted early and were using later-vintage tractors in 1940, suggesting that barley growers had high demand for tractors – a result which is likely explained by barley's exceptionally short harvest season. According to the USDA (1997, 2010), the active harvest season for barley lasts between one and three weeks; in comparison, wheat's active season lasts three weeks, while corn's active season typically extends beyond a month. The barley harvest is thus brief and frenzied, and any labor, feed, or animal shortage or mechanical failure could cause a ruinous disruption. The tractor reduced barley farms' reliance on local labor and draft animals, and the reliability of newer models was especially valuable on these farms.

The other surprising result is that rye was associated with the lowest rates of tractor adoption of all the crops studied. As Table III.1 shows, rye acreage on average generally comprises less than one percent of county farmland, so to some degree it is surprising that the coefficients on rye for tractor diffusion are estimated to be anything other than zero. But this pattern is surprising for two other reasons: counties growing other small grains adopted tractors at breakneck pace, and rye has an even shorter, more hectic harvest than barley. Further reading suggests that technical factors may have limited the effectiveness of tractors in the production of rye. Rye grows on especially large stalks, sometimes reaching over six feet in height, and when threshers or combines are fed the entire stalk, they can get jammed by the excess chaff. Another possibility is that rye shatters (breaks from the hull) easily, and the implements of the 1920s and 1930s may have been too destructive to make the tractor a sensible investment. Rye's tendency to lodge (bend, or flatten) in the field may have also limited the tractor's effectiveness for harvest.

### Robustness Checks on Baseline Results

The baseline regressions of Tables III.3 and III.4 define diffusion as the fraction of farms in a county reporting a tractor. This definition imposes an assumption of perfect indivisibility, despite evidence to the contrary of cooperative ownership (Myers 1921) and custom work (Gilbert 1930). To ensure that the results are not sensitive to this assumption, in Appendix III.C I re-estimate

[13] The vintage shares are taken over all farms answering the vintage question on the Census form, which is generally less than but close to the number of farms reporting tractors. The median rate of underreporting is 1.55 percent of farms with tractors (90th percentile: 7.83 percent of farms with tractors).

these regressions defining diffusion as the number of tractors per 100 acres of county farmland. The results of this exercise are qualitatively similar to those of the baseline regressions.

In Appendix III.D I explore the robustness of these results to weighting observations by farm count, under the premise that diffusion is the aggregation of individual adoption choices. Appendix Tables III.D.1 to III.D.4 re-estimate the specifications of Tables III.3 to III.6 applying these weights and find the results unchanged. In Appendix III.E I relax the specifications to allow for spatial correlation in the errors that is declining in the distance between county centroids up to 25-, 50-, and 100-mile cutoffs (Conley 1999). While standard errors generally increase when spatial correlation is permitted to enter the model, the basic patterns and coefficients remain highly statistically significant.

Finally, in Appendix III.F, I explore alternative explanations for these patterns such as differential New Deal relief or Dust Bowl soil erosion. To rule out the possibility that New Deal relief explains the sharp increase in diffusion in the Corn Belt, I control for county-level grant aid under the Agricultural Adjustment Act and farm credit extended under the Farm Credit Act. The effects of New Deal relief hardly register relative to those of crop mix. I also consider the possibility that the Dust Bowl may have disproportionately affected wheat-growing regions in Plains states and caused a regional shift in tractor purchases, though evidence from Hornbeck (2012, Table 1) suggests that Corn Belt counties suffered equally if not more than Wheat Belt counties. Indeed, accounting for cumulative erosion does not affect the core results.

### Counterfactual Projections

Though this evidence establishes that tractors were indeed slower to reach the Corn Belt, the magnitude of the effect is more difficult to interpret. The IV results suggest that relative to state averages, county-level diffusion in 1930 tended to be 5 percentage points (p.p.) higher for every 10 percent of farmland devoted to wheat but did not vary with the fraction of farmland in corn. In 1940, diffusion was still roughly 5 p.p. higher for every 10 percent of farmland in wheat, but it was also nearly 4 p.p. greater for every 10 percent of farmland in corn, having caught up over the intervening decade. What would aggregate diffusion in the Midwest have looked like had tractors diffused at the rate for wheat to counties more heavily concentrated in corn?

To get a better handle on this question, I use the OLS and IV estimates from the diffusion regressions to project counterfactuals. Although these estimates are linear approximations, they can offer a sense of the magnitude of the effect. Table III.7 provides actual tractor diffusion along with counterfactual projections using the OLS and IV results for both 1930 and 1940. The estimates for 1930 are calculated from the regression sample of 1,034 counties, while the estimates for 1940 are calculated from the regression sample of 941 counties. The IV-based projections suggest that tractor diffusion would have been over 25 percent higher in 1930 had tractors diffused at the same rate to the Corn Belt as they did to the Wheat Belt.

[Table III.7 about here]

Figures III.4 and III.5 disaggregate these results, plotting state-level diffusion in 1930 and 1940 for the IV counterfactuals. Not surprisingly, much of the 1930 effect comes from the Corn Belt heartland of Illinois, Iowa, and Nebraska, though a substantial amount also comes from parts of

Indiana, South Dakota, and Kansas along the Corn Belt's periphery. According to these estimates, diffusion would have been around 40 percent and roughly constant throughout the Corn Belt and Plains had it had equal scope for both wheat and corn. The tractor's technological limitations can thus explain nearly all of the cross-sectional variation in its diffusion to the major grain-producing U.S. states. By 1940, these states are observed to have similar levels of tractor diffusion, and the estimates suggest little difference between actual and counterfactual values.

[Figures III.4 and III.5 about here]

### Tractor Diffusion in Other Regions

The increasing scope of tractor diffusion is similarly apparent in most other parts of the country at this time. Table III.8 shows the fraction of farms in each state with tractors from 1920 to 2002. Between 1920 and 1940, the Midwest Census Region led the country in tractor adoption, with 42.4 percent of farms adopting by 1940. Northeastern and Western states were also mechanizing at high rates (29.2 and 27.9 percent of farms, respectively), reflecting the general utility of tractors across applications. The striking exception to this trend was the South: in states where agriculture was sharecropped, less than five percent of farms owned a tractor in 1940.

[Table III.8 about here]

Since the tractor was by this time suitable for use in cotton production, the rejection of mechanization in the South complicates the theoretical assertion that diffusion moves in lock-step with technological scope. How can this evidence be reconciled with that from other regions such as the Midwest, where generalization and diffusion seem to go hand in hand?

Researchers have largely converged on two explanations: southern agricultural labor institutions, and the difficulty of designing an affordable, functional mechanical cotton picker. The challenge of the engineering problem is summarized by Fite (1980), who catalogs the many reasons why "the nature of the cotton plant made the invention of a successful harvesting machine especially difficult." Whatley (1985, 1987) then explains how this obstacle, in conjunction with southern agrarian institutions, inhibited even partial mechanization of cotton production. Without a mechanical cotton harvester, southern farmers required a large population of farm workers to collect the harvest. In states where this labor was supplied by a migrant workforce, such as those on the Mexican border, cotton farms could mechanize pre-harvest operations without cutting into the harvest labor supply – but in most southern states, labor was supplied by annual contract in the form of tenancy and sharecropping, which was necessary to guarantee the availability of farm hands when they were needed most. Under these circumstances, mechanization became an all-or-nothing proposition: as long as the harvest technology was labor-intensive, and labor could only be secured with annual contracts, year-round operations tended to remain labor-intensive as well.

The Institutional Hypothesis (Whatley 1985) is supported by the evidence in Table III.8. In the cotton-heavy states of Texas and Oklahoma, where migrant labor was abundant, tractor diffusion increased between 1930 and 1940 to a level comparable to much of the rest of the country. According to Fite (1950), cotton mechanization in these states indeed began in the mid-1920s with the G-P tractor, and with the subsequent development of implements for pre-harvest operations, “remarkable progress occurred in the mechanization of cotton” even prior to the invention of a functional cotton picker. Given that cotton farms in these states were using tractors, the lagging diffusion in the rest of the South cannot be explained by the technology alone, which was common to both regions. The Institutional Hypothesis is further bolstered by evidence from Hornbeck and Naidu (2013), who find that farms in the Mississippi Delta mechanized in response to the large rural outmigration caused by the Great Mississippi Flood of 1927.

## III.5 Extending to Other Examples

The tractor is not unique in its history of expanding scope. In this section I show that scope was an equally important property of steam engines, electric power, and even the canonical example of hybrid corn. Each of these technologies was first developed for specific applications, with their technological limitations bounding the scope of their diffusion, and only when they generalized did they become truly pervasive. Because state-level data on hybrid acreage are available at annual frequency, hybrid corn also presents an opportunity to directly test the hypothesis that diffusion follows an S-curve both within and across distinct applications.

### Hybrid Seed Corn

While experiments with hybrid seed corn were underway at agricultural experiment stations and by private seed companies in the early twentieth century, commercial use did not meaningfully begin until the 1930s. In 1933, only around 0.1 percent of corn acreage in the U.S. was planted with hybrid seed – but by 1945, the core Corn Belt states (Illinois, Indiana, and Iowa) were nearly fully planted in hybrids, and by 1960, so was the rest of the country (Sutch 2011, 2014). Diffusion was slower to take root in regions more distant from the heart of the Corn Belt, but similarly, once it did, it was overwhelmingly swift and complete (Griliches 1960).

Griliches (1957) used the example of hybrid corn to demonstrate the basic empirical fact that technology diffusion proceeds in an S-shaped pattern over time and can be approximated by a logistic function. The argument was strictly one of scale: within a given state, crop reporting district, or the country as a whole, diffusion begins slowly, accelerates to an inflection point, and subsequently decelerates and asymptotes at its ceiling, as in Figure 1 of Griliches (1957), which I reproduce below as Figure III.6 of the present paper. When modeled with a logistic function, each of these curves can be fully characterized by three parameters: the origin, the rate of growth, and the ceiling. Griliches contends that cross-sectional variation in diffusion is driven by differences in both the rate of acceptance and the availability of locally-adapted seed varieties.

[Figure III.6 about here]

Though Griliches acknowledged availability as an important source of variation in diffusion, the subsequent literature has tended to overlook this margin. What Griliches seemingly failed to discern is that hybrid corn diffusion also proceeded along an S-curve on this extensive margin, as hybrid seed varieties were transferred or adapted to the conditions of other regions peripheral to the Corn Belt. Figure III.7 shows the distribution of states by the year they achieve each of four different levels of diffusion: 5, 10, 25, and 50 percent of acres planted. At all levels, and especially at lower levels indicating availability, the scope of diffusion follows a curve which is well-approximated by the logistic function. This evidence aligns directly with the theoretical predictions from the first section of this paper. Perhaps most strikingly, the time it took to go from a positive level of diffusion in one state to positive levels of diffusion in all 48 contiguous states exceeds the time it took to go from 0 to 100 percent diffusion within many of the individual states in this sample – a fact which speaks to the crucial role of scope in explaining lags in aggregate diffusion.

[Figure III.7 about here]

### The Steam Engine

The first practical steam engine was invented in 1712 for the exclusive purpose of pumping water out of coal mines, and for the next 60 years, steam engines were manufactured and used to pump water and little else. Not until 1781 did the technology develop to a point where it could be used more broadly: James Watt, who had previously improved on the Newcomen design by making it more powerful and fuel efficient, developed a steam engine that could produce continuous rotary motion using a crank and flywheel and power machinery with a drive belt – opening up the steam engine to effectively all industrial applications. The following year Watt developed the double-acting cylinder, which could generate both forwards and backwards belt movement. In a matter of years, Watt had thus taken the steam engine from a device with a single use to a technology that would eventually be used to power factories, vessels, and locomotives.

The steam engine fits neatly into the predictions of the theoretical model, with the exception that Watt's patent claims delayed its broader diffusion by several decades. But the history makes clear that the key to understanding the steam engine's diffusion lies in its scope. To bring this point into empirical focus, I examine the transition to steam power in U.S. manufacturing using the Atack, Bateman, and Weiss (1999; hereafter ABW) nationally representative samples from 1850 to 1870 U.S. Census of Manufactures. The ABW dataset includes a "power type" variable that specifies the principal source of power for each firm. This variable can be coded as steam, water, hand (manual), animal (draft), a combination of the above, or not given, and it enables me to trace steam power's diffusion within and across manufacturing industries. In the figures and tables that follow, I drop firms with no source given to focus on those with known power use.

For context, Table III.9 shows the share of firms reporting each power type, by decade. Manual labor was clearly the dominant source of power at this time, but steam power was on the rise.[14] Figure III.8 collapses the sample to the industry level and shows the distribution of industries in terms of the fraction of firms associating with each of manual, water, and steam power, by decade. In 1850, steam power was relatively concentrated: most industries had zero firms reporting steam power, and in industries where steam was in use, usually less than half of firms were converted. Over the following decades, steam power diffused considerably in scope, but only modestly in scale: steam power was in use in nearly three-quarters of all industries by 1870, but industries where steam power was used still tended to have less than half of firms on steam. The importance of the scope of diffusion is even more apparent when comparing steam to water, which was concentrated in just a handful of industries, thus limiting its aggregate industrial impact.

[14] The table contains some odd patterns in steam and manual power usage, most notably that the share of firms reporting manual power drops sharply in 1860 and then returns to its previous level in 1870. It is unclear why this occurs. Potential explanations include sampling variance, non-representative samples, or dislocations during the Civil War. These patterns persist if the sample is restricted to Union states east of the Mississippi and north of the Ohio Rivers, where a majority of U.S. manufacturing activity was concentrated at the time.

[Table III.9 and Figure III.8 about here]

## Electric Power

A similar exercise can be performed for electric power in U.S. manufacturing. The earliest practical dynamos were invented in 1867, and shortly after, Zenobe Gramme introduced a variant capable of providing a constant flow of high-voltage, direct current power to attached motors (which was later discovered to be a motor itself when supplied with electrical current). As a generator, the Gramme dynamo found use in electroplating, however as a motor it was limited by its inability to produce constant-speed rotary motion under variable loads, making it inadequate for other industrial uses. The first motor capable of operating at a constant speed with variable loads was invented by Frank Sprague in 1886, and together with Edison's development of power stations it marked the beginning of electrification. Using this motor, Sprague developed the first electric streetcars in Richmond, electric subway trains in Chicago, and electric elevators in New York.

Around the same time, other inventors were experimenting with alternating current, which was preferred to direct current for long-distance transmission because it reduced power loss. In the late 1880s, these inventors developed the three-phase electric generator, transformer, and induction motor that are still widely used today. The transmission advantages of alternating current led to its victory in the standards battles of the nineteenth century, and the three-phase electrical system became the basis for contemporary industrial applications and the power grid.

The 1910 Census of Manufactures provides the earliest evidence of electrification in U.S. manufacturing. The Census final report includes a table listing the primary horsepower from electricity, gas engines, steam engines, and water wheels in 102 manufacturing industries from 1899 to 1909. I use these data to trace the growing scope of electric power in manufacturing.

Table III.10 shows the share of total horsepower in U.S. manufacturing from each of the four primary power sources by year. In 1899, roughly 80 percent of manufacturing horsepower was generated by steam engines, declining to 65 percent by 1909. The share produced by water wheels fell even more precipitously, from almost 10 percent to less than 5 percent. The difference was entirely made up by electric power. In contrast, gas engines had almost no presence in manufacturing, despite contemporary technological advances and an increasing use in transportation.

[Table III.10 about here]

Figure III.9 shows the distribution of manufacturing industries by the fraction of horsepower generated from each power source, for each year in the sample. In 1899, over 40 percent of industries were not at all electrified, and of those that were, most obtained only a small fraction of their power from electricity. The industries drawing the most horsepower in the form of electricity at this time were $C$ars and general shop construction and repairs by street-railroad companies (at 59 percent) and $C$opper, tin, and sheet-iron products (41 percent), reflecting its earliest uses. By 1909, effectively all industries were partially electrified – though the streetcar industry was still the heaviest user, now drawing 87 percent of its power from electricity. Electric power thus began to take root in U.S. manufacturing over this 10 year period, diffusing both in scale and in scope, the latter supported in particular by the development of alternating current technology.[15]

[15] As the theory would suggest, the growth of electric power was likely also a result of complementary innovation, especially in electric generation and transmission, as well as changes in the production practices.

[Figure III.9 about here]

## III.6 Conclusion

As Griliches (1957) demonstrates in his seminal study of hybrid corn, technology diffusion is the consequence of both the availability of user-friendly varieties and the subsequent rate of acceptance. This paper studies the role of scope in the diffusion of farm tractors and other historically important innovations in U.S. agriculture and manufacturing. To better understand the economic underpinnings of the scope of diffusion, I begin by developing a model of innovation in specific- and general-purpose product varieties. This model suggests that new products will first develop for applications with exogenously high demand or low R&D costs. Only when the gains to specialization are exhausted will product development proceed to general-purpose variants – and only if it is technically feasible. Diffusion will in turn be shaped by the set of applications for which the product can be used. Since diffusion in scope must precede diffusion in scale, this margin can be fundamental to explaining lags in diffusion both in cross-section and in the aggregate. Moreover, due to horizontal spillovers resulting from complementarities with other innovations, inventing firms will typically have less than the socially-optimal incentives to generalize their technology for wider use, suggesting that generality should be a target for R&D policy interventions.

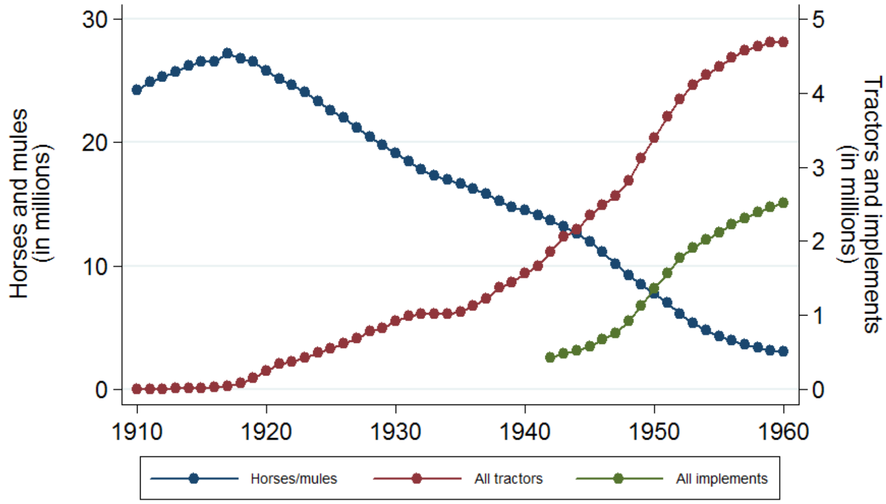
Though tractors are pervasive in modern agriculture, they were not born to be that way: the earliest models were first developed for tillage and harvesting small grains, and only in the late 1920s did the technology begin to generalize for use with row crops such as corn, cotton, and vegetables. Using county-level data on tractor ownership, crop acreage, and other variables from the 1910 to 1940 Census of Agriculture and related sources, I show that tractors were consequently quick to diffuse to areas of the U.S. Midwest growing wheat and other small grains and slower to penetrate the Corn Belt. Had the tractor diffused at the same rate to counties with equal concentrations of wheat and corn, total diffusion in the Midwest would have been on the order of 25 percent higher by 1930, and cross-sectional variation in diffusion across the major grain-producing states of the country would have been all but eliminated. Conversely, had the tractor not generalized, its impact would be so limited that it would most likely be an afterthought today.

The importance of this extensive margin of diffusion is not unique to tractors: several historically important innovations share a similar history, and many others never generalize at all. Steam engines, electric motors, and hybrid corn were each invented for applications in specific industries or geographic areas; early diffusion was accordingly limited in scope, and only when they grew more general did they become truly pervasive. Using annual data, I find that hybrid corn diffusion followed an S-shaped pattern in both its intensive and extensive margin, as predicted by the theory, and that lags in scope could be as large as or larger than those in scale.

This evidence supports a substantially different interpretation of lagging technology diffusion than what is typically found in the literature: in the examples above, lags result from a fundamental mismatch between the technology's capabilities and the technical requirements of users in different settings. The key point of departure is the idea that a single technology, at single point in time, may be used across multiple, different settings, with varying quality for each. While Suri (2011) makes the related point that the costs and benefits of technology adoption can vary across users, this heterogeneity is attributable to the quality of local infrastructure and crop yields rather than the fitness of the hybrid corn technology that is the subject of the paper per se.

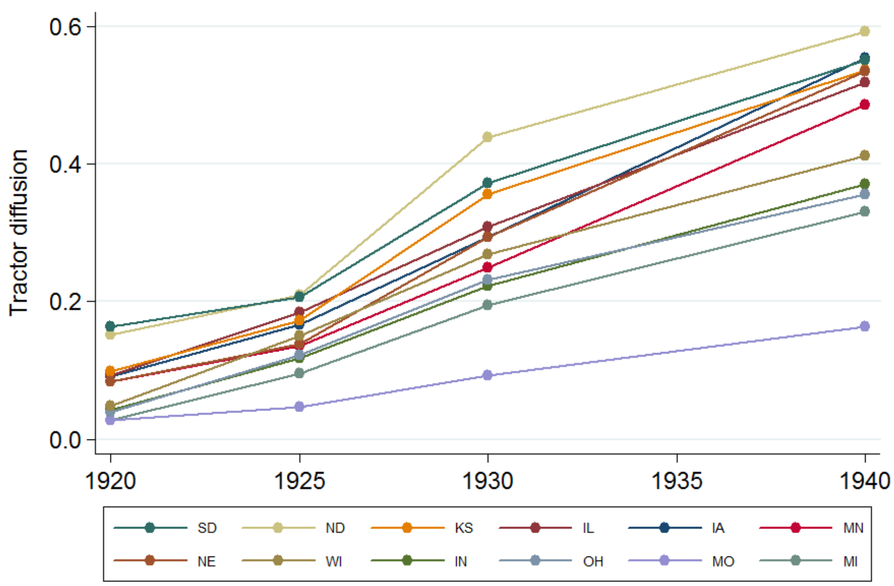
Other existing research similarly treats diffusion as varying only in scale. Early studies argued that factor prices or the fixed cost of acquiring an indivisible technology could explain lags in diffusion. More recently, researchers have turned their attention to externalities and market failures impeding technology adoption, with evidence that imperfect credit markets; assorted forms of learning and education; and present bias may all affect individual adoption decisions. But in the case of the tractor, the late-adopting U.S. Corn Belt had to wait for the row-crop tractor to be invented before farms growing corn for harvest could be fully mechanized. The results of this paper thus highlight the importance of product designs that meet the heterogeneous requirements of users in different settings, and they suggest that the most effective way to get technology into the hands of new users may simply be to make a variant adapted to their needs.

Figure III.2: Draft animals, tractors, and implements in the U.S.



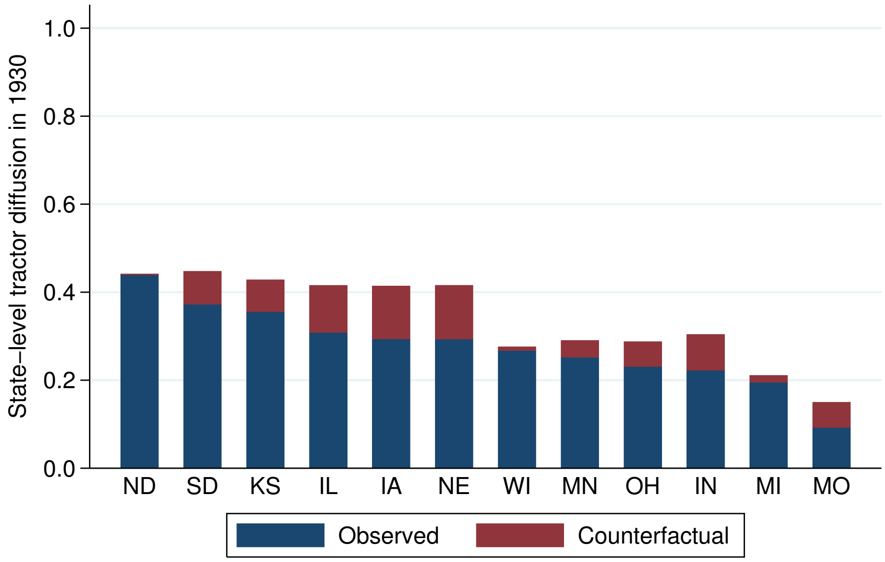

Notes: All implements refers to the sum of grain combines, corn harvesters, and pick-up hay balers owned by U.S. farms; this total does not include other implements not provided in the Historical Statistics or recorded in historical Censuses. Correlation of tractors and implements on U.S. farms is 0.996 over the 19 years for which data on all three implements are available. Data from Historical Statistics of the U.S., Series Da623, Da629-631, Da983, Da985, Da987.

Figure III.3: Tractor diffusion, by state and year



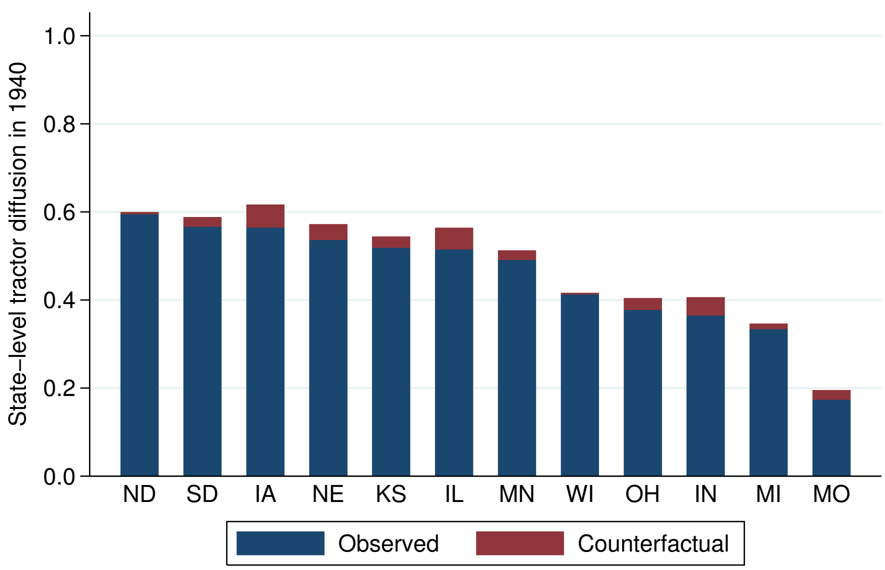

Notes: Figure shows the path of tractor diffusion for each of the twelve Midwestern states from 1920 to 1940. Data from 1940 Census of Agriculture, State Table 11.

Figure III.4: Observed and counterfactual tractor diffusion, by state, 1930 (IV estimates)



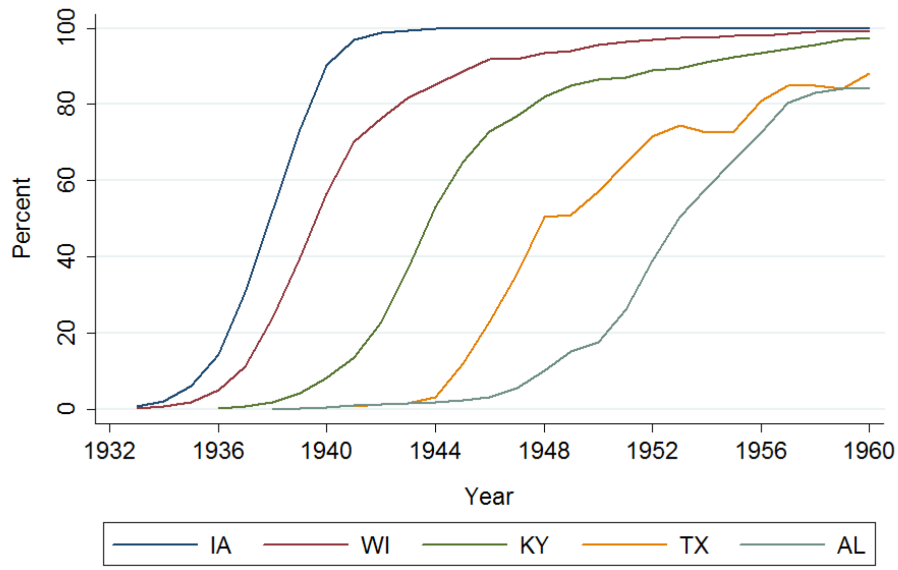

Notes: Figure shows the projected increase in tractor diffusion across all 12 states in 1930 had the tractor diffused at the same rate to counties with the same concentration of wheat and corn crops. Counterfactual calculated as described in the text.

Figure III.5: Observed and counterfactual tractor diffusion, by state, 1940 (IV estimates)



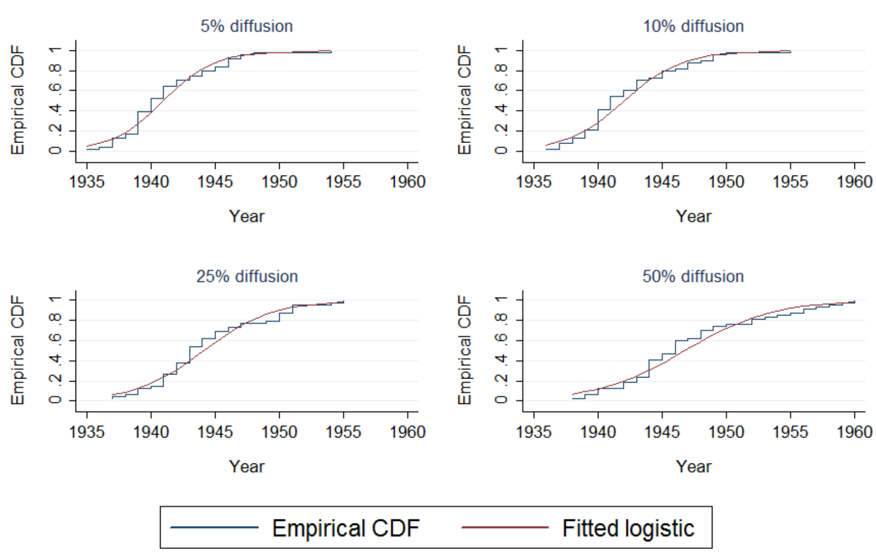

Notes: Figure shows the projected increase in tractor diffusion across all 12 states in 1940 had the tractor diffused at the same rate to counties with the same concentration of wheat and corn crops. Counterfactual calculated as described in the text.

Figure III.6: Reproduction of Griliches (1957) Fig. 1: Percentage of corn acreage planted to hybrids



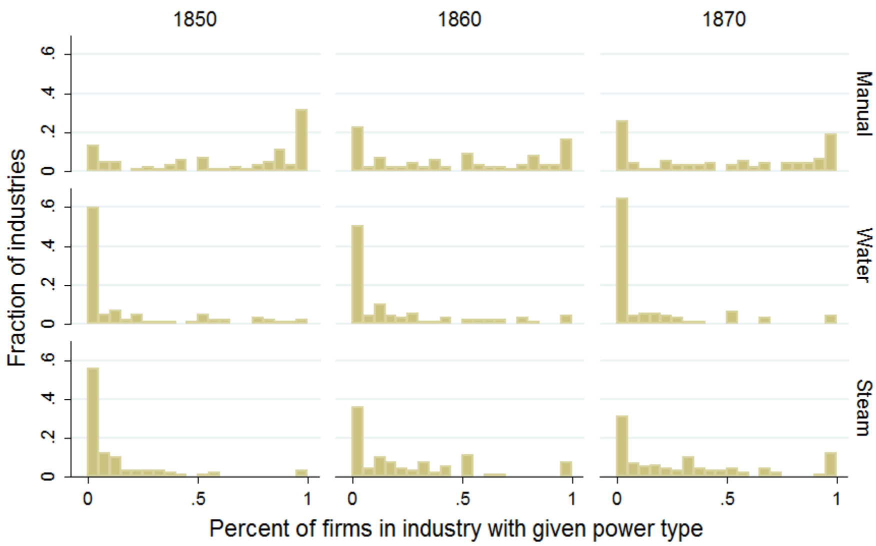

Notes: Figure shows the characteristic S-shaped hybrid corn diffusion curve for each of Iowa, Wisconsin, Kentucky, Texas, and Alabama, reproducing Figure 1 of Griliches (1957). Data from USDA Agricultural Statistics, assorted volumes.

Figure III.7: Distribution of states, by year at which given level of hybric corn diffusion attained



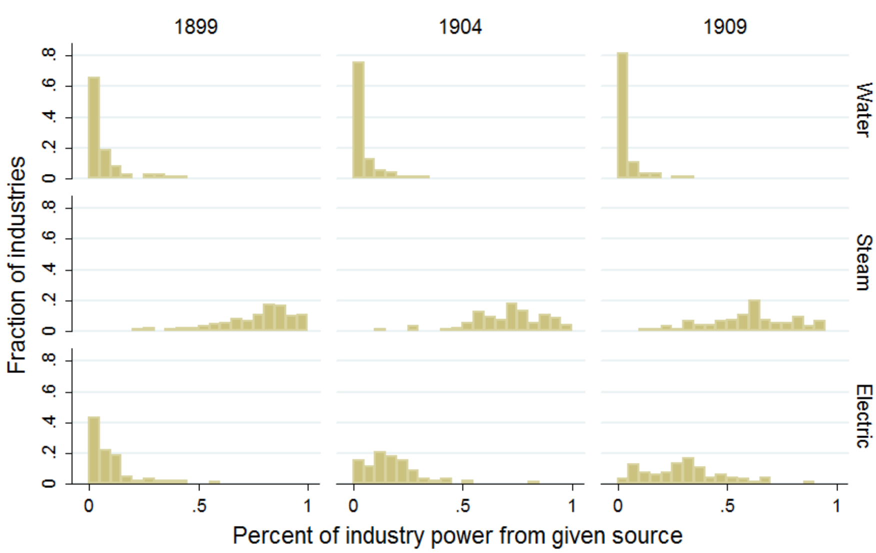

Notes: Figure shows the distribution of U.S. states by the year at which they attain a given level of hybrid corn diffusion, measured as the percentage of corn acreage planted to hybrids. Data from USDA Agricultural Statistics, assorted volumes.

Figure III.8: Distribution of manufacturing firms by source of power, 1850-1870




Notes: Figure shows distribution of U.S. manufacturing industries in the fraction of firms reporting each power type as a principal source of power, by year. Source: Atack, Bateman, and Weiss (1999) nationally representative samples from contemporary Census enumerator manuscripts.

Figure III.9: Distribution of manufacturing industries by source of power, 1899-1909




Notes: Figure shows distribution of U.S. manufacturing industries in their use of each power source as a share of all horsepower, by year, excluding industries with the five percent largest power demands in each year. Source: 1910 U.S. Census, Volume VIII (Manufactures), Chapter III, Table 3.

Table III.1: County crop mix and tractor diffusion, by year

| | **1910 (N=1035)** | **1920 (N=1035)** | **1930 (N=1035)** | **1940 (N=1035)** |
|---|---|---|---|---|
| Percent in corn | 0.168 | 0.143 | 0.145 | 0.128 |
| | (0.12) | (0.11) | (0.12) | (0.10) |
| Percent in wheat | 0.080 | 0.122 | 0.082 | 0.070 |
| | (0.09) | (0.10) | (0.11) | (0.08) |
| Percent in oats | 0.075 | 0.081 | 0.078 | 0.064 |
| | (0.06) | (0.06) | (0.07) | (0.06) |
| Percent in barley | 0.014 | 0.011 | 0.021 | 0.019 |
| | (0.03) | (0.02) | (0.03) | (0.03) |
| Percent in rye | 0.005 | 0.015 | 0.005 | 0.007 |
| | (0.01) | (0.02) | (0.01) | (0.01) |
| Percent in hay | 0.120 | 0.149 | 0.110 | 0.097 |
| | (0.04) | (0.06) | (0.05) | (0.05) |
| Tractor diffusion | | | 0.267 | 0.437 |
| | | | (0.16) | (0.21) |

Notes: Table reports average crop percentages and tractor diffusion for all counties in the sample. Crop percentages are calculated as harvested crop acreage as a fraction of county farmland. Tractor diffusion is the fraction of farms reporting tractors, available in 1930 and 1940. Standard deviations are provided in parentheses below each average.

Table III.2: Tractor diffusion, by crop intensity and year

| | **1930 (N=1035)** | | | | **1940 (N=1035)** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Above median** | **Below median** | **t-stat** | | **Above median** | **Below median** | **t-stat** | |
| Tractor diffusion, when 1910 corn acreage is: | 0.270 | 0.264 | 0.60 | | 0.473 | 0.400 | 5.77 | *** |
| | (0.13) | (0.18) | | | (0.19) | (0.22) | | |
| Tractor diffusion, when 1910 wheat acreage is: | 0.299 | 0.236 | 6.68 | *** | 0.473 | 0.400 | 5.72 | *** |
| | (0.16) | (0.15) | | | (0.19) | (0.22) | | |
| Tractor diffusion, when 1910 barley acreage is: | 0.325 | 0.210 | 12.80 | *** | 0.522 | 0.352 | 14.51 | *** |
| | (0.15) | (0.15) | | | (0.17) | (0.21) | | |

Notes: Table reports average tractor diffusion in 1930 and 1940 in counties with more and less than the median fraction of farmland in the given crop. Groups are formed based on crop acreage reported in the 1910 Census of Agriculture to avoid classifications on the basis of endogenous crop choice; at this time, tractors were owned only about 1,000 farms in the U.S. (Historical Statistics, 2006). Standard deviations are provided in parentheses below each average, and t-statistics for the difference in means is reported to the right. ***, **, and * indicate significance at the 1%, 5%, and 10% levels respectively. The table suggests that tractor adoption was not sensitive to the fraction of farmland in corn prior to the 1930s but was sensitive to the fraction of farmland in wheat and especially sensitive to the fraction of farmland in barley.

Table III.3: Effect of crop mix on tractor diffusion, 1930 and 1940; OLS

| | Diffusion | | Diff. Change | Diffusion | | Diff. Change |
|---|---|---|---|---|---|---|
| | 1930 | 1940 | 1930-1940 | 1930 | 1940 | 1930-1940 |
| Pct. in corn | -0.095* | 0.466*** | 0.296*** | 0.033 | 0.395*** | 0.164*** |
| | (0.049) | (0.070) | (0.035) | (0.059) | (0.051) | (0.043) |
| Pct. in wheat | 0.907*** | 0.788*** | 0.122*** | 0.695*** | 0.477*** | -0.019 |
| | (0.042) | (0.070) | (0.037) | (0.043) | (0.046) | (0.046) |
| Pct. in oats | 1.108*** | 1.378*** | 0.611*** | 0.995*** | 1.129*** | 0.466*** |
| | (0.069) | (0.099) | (0.053) | (0.082) | (0.070) | (0.057) |
| Pct. in barley | 0.898*** | 1.430*** | 0.728*** | 0.795*** | 1.141*** | 0.657*** |
| | (0.182) | (0.146) | (0.089) | (0.152) | (0.135) | (0.093) |
| Pct. in rye | -1.068*** | -1.465*** | 0.321** | -1.270*** | -1.166*** | 0.091 |
| | (0.309) | (0.324) | (0.141) | (0.247) | (0.247) | (0.136) |
| Constant | 0.136*** | 0.226*** | 0.077*** | 0.957** | 1.818*** | -0.702** |
| | (0.013) | (0.020) | (0.009) | (0.479) | (0.489) | (0.353) |
| N | 1034 | 954 | 954 | 1034 | 941 | 941 |
| $R^2$ | 0.70 | 0.71 | 0.67 | 0.79 | 0.90 | 0.73 |
| RMSE | 0.09 | 0.11 | 0.05 | 0.07 | 0.06 | 0.05 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor in 1930, 1940, and from 1930-1940. All specifications regress the fraction of farms with tractors on contemporaneous crop mixes. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.4: Effect of crop mix on tractor diffusion, 1930 and 1940; IV

| | Diffusion | | Diff. Change | Diffusion | | Diff. Change |
|---|---|---|---|---|---|---|
| | 1930 | 1940 | 1930-1940 | 1930 | 1940 | 1930-1940 |
| Pct. in corn | -0.271*** | 0.034 | 0.333*** | 0.056 | 0.375*** | 0.397*** |
| | (0.073) | (0.112) | (0.046) | (0.076) | (0.083) | (0.074) |
| Pct. in wheat | 0.533*** | 0.933*** | 0.229*** | 0.483*** | 0.568*** | 0.184** |
| | (0.069) | (0.118) | (0.057) | (0.062) | (0.087) | (0.079) |
| Pct. in oats | 1.394*** | 2.529*** | 0.624*** | 1.007*** | 1.357*** | 0.182 |
| | (0.105) | (0.187) | (0.095) | (0.100) | (0.123) | (0.118) |
| Pct. in barley | 0.243 | 0.589 | 0.918*** | 0.594** | 0.961*** | 1.275*** |
| | (0.313) | (0.421) | (0.209) | (0.287) | (0.358) | (0.288) |
| Pct. in rye | -1.005** | -0.953 | 0.023 | -2.452*** | -1.913*** | -0.437 |
| | (0.447) | (0.592) | (0.332) | (0.463) | (0.475) | (0.282) |
| Constant | 0.181*** | 0.218*** | 0.061*** | 1.102** | 2.095*** | -1.174** |
| | (0.015) | (0.024) | (0.011) | (0.488) | (0.646) | (0.532) |
| N | 1034 | 954 | 954 | 1034 | 941 | 941 |
| $R^2$ | 0.66 | 0.66 | 0.66 | 0.78 | 0.89 | 0.69 |
| RMSE | 0.09 | 0.12 | 0.06 | 0.07 | 0.06 | 0.05 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |
| Min. F-stat | 18.67 | 20.79 | 20.79 | 23.35 | 15.98 | 15.98 |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor in 1930, 1940, and from 1930-1940. All specifications regress the fraction of farms with tractors on contemporaneous crop mixes instrumented with pre-tractor era values. The lowest first stage F-stat is provided. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.5: Effect of crop mix on tractor vintage, 1940; OLS

| | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 |
|---|---|---|---|---|---|---|
| Pct. in corn | -0.723*** | 0.127*** | 0.596*** | -0.468*** | 0.122*** | 0.346*** |
| | (0.058) | (0.022) | (0.053) | (0.066) | (0.026) | (0.060) |
| Pct. in wheat | -0.251*** | 0.070*** | 0.181*** | -0.024 | 0.013 | 0.011 |
| | (0.053) | (0.021) | (0.045) | (0.064) | (0.025) | (0.058) |
| Pct. in oats | -0.386*** | 0.024 | 0.362*** | -0.640*** | 0.139*** | 0.501*** |
| | (0.090) | (0.034) | (0.084) | (0.094) | (0.037) | (0.083) |
| Pct. in barley | -1.221*** | 0.484*** | 0.737*** | -1.296*** | 0.520*** | 0.775*** |
| | (0.196) | (0.068) | (0.150) | (0.188) | (0.065) | (0.150) |
| Pct. in rye | -1.177*** | -0.160 | 1.337*** | -1.097*** | -0.029 | 1.126*** |
| | (0.282) | (0.098) | (0.224) | (0.296) | (0.101) | (0.237) |
| Constant | 0.450*** | 0.167*** | 0.384*** | 1.411** | 0.568** | -0.979 |
| | (0.015) | (0.006) | (0.014) | (0.668) | (0.257) | (0.599) |
| N | 954 | 954 | 954 | 941 | 941 | 941 |
| $R^2$ | 0.65 | 0.42 | 0.62 | 0.72 | 0.53 | 0.68 |
| RMSE | 0.09 | 0.04 | 0.08 | 0.08 | 0.03 | 0.08 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |

Notes: Table shows the tendency of counties with different crop mixes to own tractors of different vintages in 1940. All specifications regress the fraction of farms with tractors whose latest model-year tractor is pre-1930, 1931-35, and 1936-40 on contemporaneous crop mixes. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.6: Effect of crop mix on tractor vintage, 1940; IV

| | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 |
|---|---|---|---|---|---|---|
| Pct. in corn | -0.856*** | 0.114*** | 0.742*** | -0.592*** | 0.153*** | 0.439*** |
| | (0.091) | (0.035) | (0.082) | (0.136) | (0.047) | (0.114) |
| Pct. in wheat | -0.212** | 0.079** | 0.134 | -0.203 | 0.095** | 0.108 |
| | (0.098) | (0.034) | (0.083) | (0.132) | (0.043) | (0.113) |
| Pct. in oats | 0.304* | -0.102 | -0.202 | 0.179 | -0.162** | -0.017 |
| | (0.181) | (0.067) | (0.159) | (0.210) | (0.078) | (0.167) |
| Pct. in barley | -3.040*** | 1.046*** | 1.994*** | -3.361*** | 1.299*** | 2.063*** |
| | (0.482) | (0.164) | (0.383) | (0.640) | (0.206) | (0.498) |
| Pct. in rye | -1.104* | -0.665*** | 1.769*** | 0.381 | -0.513** | 0.132 |
| | (0.642) | (0.234) | (0.623) | (0.697) | (0.223) | (0.599) |
| Constant | 0.428*** | 0.178*** | 0.394*** | 4.171*** | -0.481 | -2.690*** |
| | (0.021) | (0.008) | (0.019) | (1.048) | (0.369) | (0.875) |
| N | 954 | 954 | 954 | 941 | 941 | 941 |
| $R^2$ | 0.59 | 0.34 | 0.57 | 0.63 | 0.40 | 0.63 |
| RMSE | 0.10 | 0.04 | 0.09 | 0.10 | 0.04 | 0.08 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |
| Min. F-stat | 20.79 | 20.79 | 20.79 | 15.98 | 15.98 | 15.98 |

Notes: Table shows the tendency of counties with different crop mixes to own tractors of different vintages in 1940. All specifications regress the fraction of farms with tractors whose latest model-year tractor is pre-1930, 1931-35, and 1936-40 on contemporaneous crop mixes instrumented with pre-tractor era values. The lowest first stage F-stat is provided. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.7: Counterfactual Midwest tractor diffusion, 1930 and 1940

| | **1930** | (% Increase) | **1940** | (% Increase) |
|---|---|---|---|---|
| **Actual diffusion** | 0.256 | | 0.431 | |
| **OLS counterfactual** | 0.358 | 39.7% | 0.443 | 2.6% |
| **IV counterfactual** | 0.322 | 25.6% | 0.458 | 6.3% |
| **Sample size** | 1034 | | 941 | |

Notes: Table reports projections of tractor diffusion across the U.S. Midwest in 1930 and 1940 had the tractor diffused at the same rate to counties with equal concentrations of wheat and corn. The estimates in this table are approximations, projected from the OLS and IV estimates (respectively) in Tables III.3 and III.4, as described in the text. The sample for these estimates (including for the calculation of actual diffusion above) is restricted to counties that were included in the regressions.

Table III.8: Percent of farms with tractors, by region and year

| **Census Region** | **1920** | **1925** | **1930** | **1940** | **2002** |
|---|---|---|---|---|---|
| **Northeast** | 2.7 | 9.5 | 18.6 | 29.2 | 86.2 |
| **Midwest** | 6.8 | 13.6 | 25.7 | 42.4 | 89.6 |
| **South** | 1.0 | 2.3 | 4.0 | 7.9 | 91.8 |
| *excl. DE, MD, OK, TX* | 0.7 | 1.8 | 2.7 | 4.2 | 90.0 |
| *DE, MD alone* | 2.8 | 7.5 | 15.5 | 23.0 | 90.3 |
| *OK, TX alone* | 2.2 | 3.7 | 7.9 | 21.3 | 95.0 |
| **West** | 7.0 | 10.7 | 19.4 | 27.9 | 83.2 |

Notes: Table reports percent of farms in each region owning a tractor in 1920, 1925, 1930, 1940, and 2002. The table highlights the lagging adoption of tractors in Southern states through 1940, especially those with historically poor labor institutions (slavery and sharecropping), and their eventual catch-up to the rest of the country. Source: 1940 U.S. Census of Agriculture, Volume 1, State Table 11.

Table III.9: Shares of firms in U.S. manufacturing, by power source, 1850-1870

| | **Steam** | **Water** | **Draft** | **Manual** | **Combo** |
|---|---|---|---|---|---|
| **1850 (N=5179)** | 0.077 | 0.297 | 0.044 | 0.552 | 0.030 |
| **1860 (N=3975)** | 0.203 | 0.335 | 0.036 | 0.397 | 0.029 |
| **1870 (N=4439)** | 0.202 | 0.181 | 0.029 | 0.572 | 0.016 |

Notes: Table reports the share of U.S. manufacturing firms associating with each of five different power sources. Sample excludes firms not providing a primary power source or for which the primary power source is unavailable. Source: Atack, Bateman and Weiss nationally representative samples from manuscript Censuses of Manufactures, 1850-1870.

Table III.10: Shares of horsepower in U.S. manufacturing, 1899-1909

| | **Electric** | **Gas** | **Steam** | **Water** |
|---|---|---|---|---|
| **1899 (N=93)** | 0.065 | 0.044 | 0.800 | 0.091 |
| **1904 (N=97)** | 0.147 | 0.061 | 0.720 | 0.072 |
| **1909 (N=97)** | 0.247 | 0.052 | 0.653 | 0.048 |

Notes: Table reports the share of horsepower in U.S. manufacturing held by the four principal power sources in industry at the turn of the 20th century. Horsepower shares are aggregated over all industries reported in the 1910 U.S. Census of Manufactures, censoring industries with the 5% largest power demands in each year. Source: 1910 U.S. Census, Volume VIII (Manufactures), Chapter III, Table 3.

# Bibliography

**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.

**Akcigit, Ufuk and Qingmin Liu**, "The Role of Information in Innovation and Competition," 2014. Working Paper.

**Alston, Lee J.**, "Farm Foreclosures in the United States during the Interwar Period," *The Journal of Economic History*, 1983, *43* (4), 885–903.

**Amabile, Teresa M. and Mukti Khaire**, "Creativity and the Role of the Leader," *Harvard Business Review*, 2008, *October*, 100–109.

**Ankli, Robert and Alan Olmstead**, "The Adoption of the Gasoline Tractor in California," *Agricultural History*, 1981, *55* (3), 213–230.

**Aoyagi, Masaki**, "Information Feedback in a Dynamic Tournament," *Games and Economic Behavior*, 2010, *70* (2), 242–260.

**Arrow, Kenneth J.**, "Economic Welfare and the Allocation of Resources for Invention," in "The Rate and Direction of Inventive Activity: Economic and Social Factors," Princeton: Princeton University Press, 1962.

**Atack, Jeremy, Fred Bateman, and Thomas Weiss**, *Samples from the Manuscript Censuses of Manufactures, 1850-1880.* Available at: https://my.vanderbilt.edu/jeremyatack/data-downloads/.

**Azmat, Ghazala and Nagore Iriberri**, "The Importance of Relative Performance Feedback Information: Evidence from a natural experiment using high school students," *Journal of Public Economics*, 2010, *94* (7), 435–452.

**Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso**, "Incentives and Creativity: Evidence from the Academic Life Sciences," *RAND Journal of Economics*, 2011, *42* (3), 527–554.

**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.

**Beer, Michael**, "Performance Appraisal," in Jay W. Lorsch, ed., *Handbook of Organizational Behavior*, Englewood Cliffs: Prentice Hall, 1987.

**Bénabou, Roland and Jean Tirole**, "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 2003, *70* (3), 489–520.

**Bergemann, Dirk and Juuso Välimäki**, "Bandit Problems," in Steven N. Durlauf and Lawrence E. Blume, eds., *The New Palgrave Dictionary of Economics*, 2 ed., Basingstoke: Palgrave Macmillan, 2008.

**Bimpikis, Kostas, Shayan Ehsani, and Mohamed Mostagir**, "Information Provision in Dynamic Innovation Tournaments," 2014. Working Paper.

**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," 2014. Research Policy (forthcoming).

**_ , _ , and Michael Menietti**, "Performance Responses To Competition Across Skill-Levels In Rank Order Tournaments: Field Evidence and Implications For Tournament Design," 2014. Working Paper, available at SSRN.

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

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

**Brunt, Liam, Josh Lerner, and Tom Nicholas**, "Inducement Prizes and Innovation," *Journal of Industrial Economics*, 2012, *60* (4), 657–696.

**Carter, Susan B., Scott Sigmund Gartner, Michael R. Haines, Alan L. Olmstead, Richard Sutch, and Gavin Wright, eds**, *Historical Statistics of the United States, Earliest Times to the Present: Millenial Edition*, New York: Cambridge University Press, 2006.

**Charness, Gary and Daniela Grieco**, "Creativity and Financial Incentives," 2014. Working Paper.

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

**Choi, Jay P.**, "Dynamic R&D competition under "hazard rate" uncertainty," *Rand Journal of Economics*, 1991, *22* (4), 596–610.

**Clarke, Sally**, "New Deal Regulation and the Revolution in American Farm Productivity: A Case Study of the Diffusion of the Tractor in the Corn Belt, 1920-1940," *The Journal of Economic History*, 1991, *51* (1), 101–123.

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

**Conley, Timothy G.**, "GMM Estimation with Cross-Sectional Dependence," *Journal of Econometrics*, 1999, *92* (1), 1–45.

**_ and Christopher R. Udry**, "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, 2010, *100* (1), 35–69.

**David, Paul A.**, "The Mechanization of Reaping in the Ante-bellum Midwest," in Henry Rosovsky, ed., *Industrialization in Two Systems: Essays in Honor of Alexander Gerschenkron*, John Wiley & Sons, 1966, pp. 3–39.

**Donaldson, Dave and Richard Hornbeck**, "Railroads and American Economic Growth: A "Market Access" Approach," 2013. NBER working paper 19213.

**Duflo, Esther, Michael Kremer, and Jonathan Robinson**, "How High are Rates of Return to Fertilizer? Evidence from Field Experiments in Kenya," *American Economic Review: Papers & Proceedings*, 2008, *98* (2), 482–488.

**—, —, and —**, "Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya.," *American Economic Review*, 2011, *101* (6), 2350–90.

**Dunning, Lorry**, *The Ultimate American Farm Tractor Data Book*, Osceola: MBI, 1999.

**Dupas, Pascaline**, "Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence From a Field Experiment," *Econometrica*, 2014, *82* (1), 197–228.

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

**— and Ernst Fehr**, "Deception and Incentives: How Dishonesty Undermines Effort Provision," 2009. Working Paper.

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

**Eisenberger, Robert and Linda Rhoades**, "Incremental Effects of Reward on Creativity," *Journal of Personality and Social Psychology*, 2001, *81* (4), 728–741.

**Eriksson, Tor, Anders Poulsen, and Marie Claire Villeval**, "Feedback and Incentives: Experimental evidence," *Labour Economics*, 2009, *16* (6), 679–688.

**Federal Deposit Insurance Corporation**, *Bank Deposit Data* 1920-1936. Available from the National Historical Geographic Information System: https://www.nhgis.org/.

**Fishback, Price V., Shawn Kantor, and John Wallis**, "Can the New Deal's Three R's Be Rehabilitated? A Program-by-Program, County-by-County Analysis," *Explorations in Economic History*, 2003, *40* (3), 278–307.

**—, William C. Horrace, and Shawn Kantor**, "Did New Deal Grant Programs Stimulate Local Economies? A Study of Federal Grants and Retail Sales During the Great Depression," *The Journal of Economic History*, 2005, *65* (1), 36–71.

**—, —, and —**, "The Impact of New Deal Expenditures on Mobility During the Great Depression," *Explorations in Economic History*, 2006, *43* (2), 179–222.

**Fite, Gilbert C.**, "Recent Progress in the Mechanization of Cotton Production in the United States," *Agricultural History*, 1950, *24* (1), 19–28.

**—**, "Mechanization of Cotton Production Since World War II," *Agricultural History*, 1980, *54* (1), 190–207.

**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.

**Gershkov, Alex and Motty Perry**, "Tournaments with Midterm Reviews," *Games and Economic Behavior*, 2009, *66* (1), 162–190.

**Gilbert, C.**, *An Economic Study of Farm Tractors in New York*, Ithaca: Cornell University, 1930.

**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.

**Gill, David**, "Strategic Disclosure of Intermediate Research Results," *Journal of Economics & Management Strategy*, 2008, *17* (3), 733–758.

**Goltsman, Maria and Arijit Mukherjee**, "Interim Performance Feedback in Multistage Tournaments: The optimality of partial disclosure," *Journal of Labor Economics*, 2011, *29* (2), 229–265.

**Green, Jerry R. and Suzanne Scotchmer**, "On the division of profit in sequential innovation," *RAND Journal of Economics*, 1995, *26* (1), 20–33.

**Griliches, Zvi**, "Hybrid Corn: An Exploration in the Economics of Technological Change," *Econometrica*, 1957, *25* (4), 501–522.

_ , "Hybrid Corn and the Economics of Innovation," *Science*, 1960, *132* (3422), 275–280.

**Gross, Daniel P.**, "Scale vs. Scope in the Diffusion of New Technology: Evidence from the Farm Tractor (and Other Examples)," 2014. Working Paper.

_ , "Creativity Under Fire: The Effects of Competition on Innovation and the Creative Process," 2015. Working Paper.

_ , "Interim Feedback in Tournaments for Ideas: Trading Off Participation for Quality," 2015. Working Paper.

**Haile, Philip A. and Elie Tamer**, "Inference with an Incomplete Model of English Auctions," *Journal of Political Economy*, 2003, *111* (1), 1–51.

**Halac, Marina, Navin Kartik, and Qingmin Liu**, "Contests for Experimentation," 2015. Working Paper.

**Hendel, Igal and Yossi Spiegel**, "Small Steps for Workers, a Giant Leap for Productivity," *American Economic Journal: Economic Policy*, 2014, *6* (1), 73–90.

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

**Hornbeck, Richard**, "Barbed Wire: Property Rights and Agricultural Development," *The Quarterly Journal of Economics*, 2010, *125* (2).

_ , "The Enduring Impact of the American Dust Bowl: Short- and Long-Run Adjustments to Environmental Catastrophe," *American Economic Review*, 2012, *102* (4), 1477–1507.

_ **and Suresh Naidu**, "When the Levee Breaks: Black Migration and Economic Development in the American South." Forthcoming in American Economic Review.

**Igami, Mitsuru**, "Estimating the Innovator's Dilemma: Structural Analysis of Creative Destruction," 2013. Working Paper.

**Iverson, G.**, "Possibilities of the All-Purpose Tractor," *Agricultural Engineering*, 1922, *3*, 147.

**Jones, Charles I.**, "R&D-Based Models of Economic Growth," *Journal of Political Economy*, 1995, *103* (4), 759–784.

_ **and John C. Williams**, "Measuring the Social Return to R&D," *Quarterly Journal of Economics*, 1998, *113* (4), 1119–1135.

_ **and** _ , "Too Much of a Good Thing? The Economics of Investment in R&D," *Journal of Economic Growth*, 2000, *5* (1), 65–85.

**Keller, Wolfgang**, "Geographic Localization of International Technology Diffusion," *American Economic Review*, 2002, *92* (1), 120–142.

**Klancher, Lee**, *The Farmall Dynasty*, St. Paul: 671 Press, 2008.

**Lazear, Edward and Sherwin Rosen**, "Rank-Order Tournaments as Optimum Labor Contracts," *Journal of Political Economy*, 1981, *89* (5), 841–864.

**Lazear, Edward P.**, "Firm-Specific Human Capital: A Skill-Weights Approach," *Journal of Political Economy*, 2009, *117* (5), 914–940.

**Leffingwell, Randy**, *Ford Farm Tractors*, Minneapolis: Voyageur Press, 1998.

_ , *The American Farm Tractor*, Minneapolis: Voyageur Press, 2002.

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

**Manuelli, Rodolfo E. and Ananth Seshadri**, "Frictionless Technology Diffusion: The Case of Tractors," *American Economic Review*, 2014, *104* (4), 1368–1391.

**McFadden, Daniel**, "The Measurement of Urban Travel Demand," *Journal of Public Economics*, 1974, *3* (4), 303–328.

**Minnesota Population Center**, *National Historical Geographic Information System: Version 2.0*, Minneapolis: University of Minnesota, 2011.

**Mokyr, Joel**, *The Lever of Riches: Technological Creativity and Economic Progress*, New York: Oxford University Press, 1990.

**Morgan, John and Richard Wang**, "Tournaments for Ideas," *California Management Review*, 2010, *52* (2), 77–97.

**Moser, Petra and Tom Nicholas**, "Prizes, Publicity and Patents: Non-monetary Awards as a Mechanism to Encourage Innovation," *Journal of Industrial Economics*, 2013, *61* (3), 763–788.

**Munshi, Kaivan**, "Social learning in a heterogeneous population: technology diffusion in the Indian Green Revolution," *Journal of Development Economics*, 2004, *73* (1), 185–213.

**Murphy, K. R. and J. N. Cleveland**, *Performance Appraisal: An Organizational Perspective*, Boston: Allyn and Bacon, 1995.

**Musoke, Moses S.**, "Mechanizing Cotton Production in the American South: The Tractor, 1915-1960," *Explorations in Economic History*, 1981, *18* (4), 347–375.

**Myers, W.**, *An Economic Study of Farm Tractors in New York*, Ithaca: Cornell University, 1921.

**Olmstead, Alan L.**, "The Mechanization of Reaping and Mowing in American Agriculture, 1833-1870," *The Journal of Economic History*, 1975, *35* (2), 327–352.

_ **and Paul W. Rhode**, "The Agricultural Mechanization Controversy of the Interwar Years," *Agricultural History*, 1994, *68* (3), 35–53.

_ **and** _ , "Reshaping the Landscape: The Impact and Diffusion of the Tractor in American Agriculture, 1910-1960," *The Journal of Economic History*, 2001, *61* (3), 663–698.

**PRISM Climate Group at Oregon State University**, *United States Average Monthly or Annual Precipitation, 1971-2000 and 1981-2010* 2012. Available at: http://prism.oregonstate.edu/products/matrix.phtml?vartype=ppt&view=data.

**Rieck, Thomas**, "Information Disclosure in Innovation Contests," 2010. Working Paper.

**Sanders, Ralph W.**, *The Farm Tractor: 100 Years of North American Tractors*, Minneapolis: Voyageur Press, 2009.

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

**Scotchmer, Suzanne**, "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law," *Journal of Economic Perspectives*, 1991, *5* (1), 29–41.

**Steckel, Richard H. and William J. White**, "Engines of Growth: Farm Tractors and Twentieth-Century US Economic Welfare." NBER Working Paper 17879.

**Stern, Scott**, "Do Scientists Pay to Be Scientists?," *Management Science*, 2004, *50* (6), 835–853.

**Suri, Tavneet**, "Selection and Comparative Advantage in Technology Adoption," *Econometrica*, 2011, *79* (1), 159–209.

**Sutch, Richard**, "The Impact of the 1936 Corn Belt Drought on American Farmers' Adoption of Hybrid Corn," in Gary D. Libecap and Richard H. Steckel, eds., *The Economics of Climate Change: Adaptations Past and Present*, Chicago: University of Chicago Press, 2011, pp. 195–223.

_ , "America's First BioTech Food: The Strangely-Political Story of Hybrid Corn," 2014. Prepared for presentation to the 2014 Roger Smith Conference on Food.

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

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

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

**Tullock, Gordon**, "Efficient Rent Seeking," in James Buchanan, Robert Tollison, and Gordon Tullock, eds., *Toward a Theory of the Rent-seeking Society*, College Station: Texas A&M University Press, 1980.

**U.S. Census Bureau**, *Thirteenth Census of the United States*, Washington: GPO, 1910.

_ , *Census of Agriculture*, Washington: GPO, 1910-1954.

_ , *Census of Agriculture*, Washington: GPO, 2002.

_ , *Census of Agriculture*, Washington: GPO, 2007.

**U.S. Dept. of Justice and Federal Trade Commission**, *Horizontal Merger Guidelines* 2010. Available at: http://www.justice.gov/atr/public/guidelines/hmg-2010.html.

**USDA**, *Usual Planting and Harvesting Dates for U.S. Field Crops* 1997. Available at: http://www.nass.usda.gov/Publications/Usual_Planting_and_Harvesting_Dates/uph97.pdf.

_ , *Usual Planting and Harvesting Dates for U.S. Field Crops* 2010. Available at: http://usda01.library.cornell.edu/usda/nass/planting/2010s/2010/planting-10-29-2010.pdf.

**USGS**, *National Elevation Dataset* 2009. Available at: http://ned.usgs.gov/.

**Weitzman, Martin**, "Optimal Search for the Best Alternative," *Econometrica*, 1979, *47* (3), 641–54.

— , "Hybridizing Growth Theory," *American Economic Review: Papers and Proceedings*, 1996, *86* (2), 207–212.

— , "Recombinant Growth," *Quarterly Journal of Economics*, 1998, *113* (2), 331–360.

**Whatley, Warren C.**, "A History of Mechanization in the Cotton South: The Institutional Hypothesis," *Quarterly Journal of Economics*, 1985, *100* (4), 1191–1215.

— , "Southern Agrarian Labor Contracts as Impediments to Cotton Mechanization," *Journal of Economic History*, 1987, *47* (1), 45–70.

**White House Office of Science and Technology Policy (OSTP)**, *Implementation of Federal Prize Authority: Fiscal Year 2013 Progress Report* 2014.

**White, William J.**, "An Unsung Hero: The Tractor's Contribution to Twentieth-Century United States Economic Growth," 2000. Unpublished dissertation.

— , "Economic History of Tractors in the United States," 2010. EH.Net Encyclopedia. Available at: http://eh.net/encyclopedia/article/white.tractors.history.us.

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

**Williams, Robert C.**, *Fordson, Farmall, and Poppin' Johnny: A History of the Farm Tractor and Its Impact on America*, Urbana: University of Illinois Press, 1987.

**Wirtz, Julia**, "Feedback and Learning in Tournaments," 2014. Working Paper.

**W.J. Hartman Publishing**, *Hartman's Western Freight Rates*, Chicago: W.J. Hartman, 1916.

**Wooten, Joel and Karl Ulrich**, "Idea Generation and the Role of Feedback: Evidence from field experiments with innovation tournaments," 2014. Working Paper, available at SSRN.

**Wright, Brian D.**, "The Economics of Innovation Incentives: Patents, Prizes, and Research Contracts," *American Economic Review*, 1983, *73* (4), 691–707.

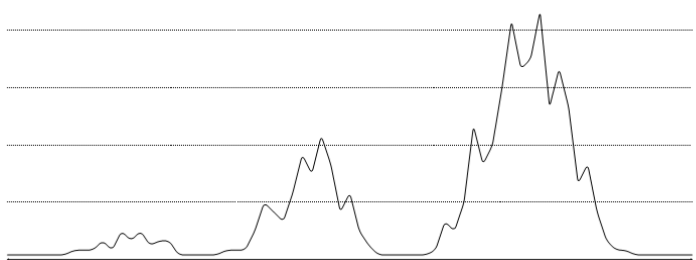
**Yildirim, Huseyin**, "Contests with Multiple Rounds," *Games and Economic Behavior*, 2005, *51* (1), 213–227.

# Appendix for Chapter I

## I.A Analogy for Search Process

A simple analogy can illustrate the nature and consequences of experimentation in this setting. Consider a sponsor with smooth preferences $f(\cdot)$ over the real line, mapped as a function over $\mathbb{R}$, as illustrated in Figure I.A.1. This function is ex-ante unknown to players, and perhaps even to the sponsor, but known to exist. Players begin by drawing a random number $x$ and learning the value of $f(x)$. Since $f(\cdot)$ is smooth, the player knows that any incremental movement along $\mathbb{R}$, left or right, to $x'$ will yield $f(x') \approx f(x)$; the outcome of a more radical deviation from $x$ is uncertain.

Figure I.A.1: Sponsor's preferences over the real line



The player can use incremental search to seek out a local maximum, or the highest local maximum in a neighborhood of $x$, but to identify even higher local maxima or a global maximum, she will need to experiment. Experimentation from a relatively favorable initial $x$ is likely to be high mean, but also high variance, as in Figure I.A.1. These are features of the model in Section I.1: the expected outcome of experimentation, $q\alpha\beta_1 + (1-q)\frac{1}{\alpha}\beta_1$, and the difference between upside and downside outcomes, $\alpha\beta_1 - \frac{1}{\alpha}\beta_1$, are both increasing in $\beta_1$, the player's initial draw. Section I.4 shows they are also features of the empirical setting.

In Figure I.A.1, the probability of successful experimentation after an initial draw of $x$ is a function of $f(x)$: the better a player's existing draw, the less likely it is that experimentation will yield an even better one (and vice versa). Intuitively, when the best draw is very good, experimentation has mostly downside; when the best draw is low, it has mostly upside. This feature can naturally be incorporated into the model of Section I.1 by endogenizing $q = q(\beta_1)$, with $q(\cdot)$ a decreasing, convex function of $\beta_1$ (a simple, parametric example is $q = \exp\{-\beta_1\}$). Doing so does not fundamentally change the results of Sections I.1 to I.1, since those results obtain from comparative statics with respect to the level of competition $\mu$. Section I.4 shows that the probability of high and low experimentation outcomes varies as expected with the initial draw.

The conclusion of Section I.1 is that competition can shape incentives for experimentation when this search process is embedded in a tournament: a player with a high-quality design may be induced to experiment by the competition she faces. This result holds as long as this probability of successful experimentation is greater than the minimum threshold derived in the paper.

## I.B Proofs of Theorems and Additional Figures

**L**emma 1: The gains to exploration over abandonment are increasing and concave in $\mu$ when $\mu$ is small and decreasing and convex when $\mu$ is large. The gains are zero when $\mu = 0$ and approach zero from above as $\mu \to \infty$, holding $\beta_1$ fixed.

**P**roof:

Part (i): Low $\mu$. As $\mu \longrightarrow 0$:

$$\begin{aligned}
qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(0) &= q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1+\mu}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1+\mu}\right)-\left(\frac{\beta_1}{\beta_1+\mu}\right)\\
&\longrightarrow q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1}\right)-\left(\frac{\beta_1}{\beta_1}\right)\\
&= q+(1-q)-1=0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\mu}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(0)\right] &= q\left(\frac{-(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^2}\right)+(1-q)\left(\frac{-\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^2}\right)+\frac{\beta_1}{(\beta_1+\mu)^2}\\
&\longrightarrow q\left(\frac{-1}{(1+\alpha)\beta_0}\right)+(1-q)\left(\frac{-1}{\left(1+\frac{1}{\alpha}\right)\beta_0}\right)+\frac{1}{\beta_0}\\
&= \frac{-q\left(1+\frac{1}{\alpha}\right)-(1-q)(1+\alpha)+(1+\alpha)\left(1+\frac{1}{\alpha}\right)}{(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_1}\\
&= \frac{-\left(q+\frac{1}{\alpha}q\right)-(1-q+\alpha-\alpha q)+\left(1+\alpha+\frac{1}{\alpha}+1\right)}{(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_1}\\
&= \frac{\alpha q-\frac{1}{\alpha}q+\alpha+\frac{1}{\alpha}}{(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_1}=\frac{\left(\alpha^2-1\right)q+\left(1+\alpha^2\right)}{(1+\alpha)^2\beta_1}\longrightarrow 0^+
\end{aligned}$$

$$\frac{\partial^2}{\partial\mu^2}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(0)\right]=q\left(\frac{2(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^3}\right)+(1-q)\left(\frac{2\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^3}\right)-\frac{2\beta_1}{(\beta_1+\mu)^3}$$

$$\text{num.}\longrightarrow\left(q\cdot(1+\alpha)\left(1+\tfrac{1}{\alpha}\right)^3+(1-q)\cdot\left(1+\tfrac{1}{\alpha}\right)(1+\alpha)^3-(1+\alpha)^3\left(1+\tfrac{1}{\alpha}\right)^3\right)2\beta_1^7<0$$

Part (ii): High $\mu$. As $\mu \longrightarrow \infty$:

$$\begin{aligned}
qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(0) &= q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1+\mu}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1+\mu}\right)-\left(\frac{\beta_1}{\beta_1+\mu}\right)\\
&\longrightarrow \frac{1}{\mu}\left(q(1+\alpha)\beta_1+(1-q)\left(1+\frac{1}{\alpha}\right)\beta_1-\beta_1\right)\\
&= \frac{1}{\mu}\beta_1\left(\alpha q-\frac{1}{\alpha}q+\frac{1}{\alpha}\right)\\
&= \frac{1}{\alpha\mu}\beta_1\left(\left(\alpha^2-1\right)q+1\right)\longrightarrow 0^+
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\mu}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(0)\right] &= q\left(\frac{-(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^2}\right)+(1-q)\left(\frac{-\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^2}\right)+\frac{\beta_1}{(\beta_1+\mu)^2}\\
&\longrightarrow \frac{1}{\mu^2}\left(-q(1+\alpha)\beta_1-(1-q)\left(1+\frac{1}{\alpha}\right)\beta_1+\beta_1\right)\\
&= \frac{1}{\mu^2}\beta_1\left(-(q+\alpha q)-\left(1-q+\frac{1}{\alpha}-\frac{1}{\alpha}q\right)+1\right)\\
&= \frac{1}{\mu^2}\beta_1\left(-\alpha q+\frac{1}{\alpha}q-\frac{1}{\alpha}\right)\\
&= \frac{1}{\alpha\mu^2}\beta_1\left(-\left(\alpha^2-1\right)q-1\right)\longrightarrow 0^-
\end{aligned}$$

$$\frac{\partial^2}{\partial\mu^2}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(0)\right]=q\left(\frac{2(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^3}\right)+(1-q)\left(\frac{2\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^3}\right)-\frac{2\beta_1}{(\beta_1+\mu)^3}$$

num. $\longrightarrow \left(q\cdot(1+\alpha)+(1-q)\cdot\left(1+\frac{1}{\alpha}\right)-1\right)2\beta_1\mu^6>0$

Taken together, these asymptotics generate a curve with the shape described.

**P**roposition 1: For all values of $q$, there exists a unique level of competition $\mu_1^*$ at which the gains to exploration, relative to abandonment, are maximized.

**P**roof: Existence follows from lemma and continuity of the success function. Since the difference of the success function under exploration and abandonment is quadratic in $\mu$, it has at most two real roots, one of which is shown below to be zero, the other of which is shown to be negative. Given the shape described by the lemma, the value at which this difference is maximized must be unique.

To find the roots, set $qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(0)=0$ and solve for $\mu$:

$$
\begin{aligned}
0 &= q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1+\mu}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1+\mu}\right)-\left(\frac{\beta_1}{\beta_1+\mu}\right)\\
&= q(1+\alpha)\beta_1\left[\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)(\beta_1+\mu)\right]\\
&\quad +(1-q)\left(1+\frac{1}{\alpha}\right)\beta_1\left[((1+\alpha)\beta_1+\mu)(\beta_1+\mu)\right]\\
&\quad -\beta_1\left[((1+\alpha)\beta_1+\mu)\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)\right]\\
&= \mu^2\beta_1\left[q(1+\alpha)+(1-q)\left(1+\frac{1}{\alpha}\right)-1\right]\\
&\quad +\mu\beta_1^2\left[q\left(2+\alpha+\frac{1}{\alpha}\right)+(1-q)\left(2+\alpha+\frac{1}{\alpha}\right)+q(1+\alpha)+(1-q)\left(1+\frac{1}{\alpha}\right)-\left(2+\alpha+\frac{1}{\alpha}\right)\right]\\
&\quad +\beta_1^3\left[q\left(2+\alpha+\frac{1}{\alpha}\right)+(1-q)\left(2+\alpha+\frac{1}{\alpha}\right)-\left(2+\alpha+\frac{1}{\alpha}\right)\right]\\
&= a\beta_1\mu^2+b\beta_1^2\mu+c\beta_1^3\,,
\end{aligned}
$$

where

$$
\begin{aligned}
a &= q+\alpha q+(1-q)+\frac{1}{\alpha}(1-q)-1=\alpha q+\frac{1}{\alpha}(1-q)=\frac{1}{\alpha}\left(\left(\alpha^2-1\right)q+1\right)\begin{cases}>0 & \text{if } q<\frac{1}{1-\alpha^2}\\ <0 & \text{if } q>\frac{1}{1-\alpha^2}\end{cases}\\
b &= q+\alpha q+(1-q)+\frac{1}{\alpha}(1-q)=\alpha q+\frac{1}{\alpha}(1-q)+1=\frac{1}{\alpha}(\alpha+1)((\alpha-1)q+1)\begin{cases}>0 & \text{if } q<\frac{1}{1-\alpha}\\ <0 & \text{if } q>\frac{1}{1-\alpha}\end{cases}\\
c &= 0
\end{aligned}
$$

By the quadratic formula, the roots are thus:

$$
\frac{-\left(b\beta_1^2\right)\pm\sqrt{\left(b\beta_1^2\right)^2-0}}{2(a\beta_1)}=\frac{-(b\beta_1)\pm-(b\beta_1)}{2a}=\begin{cases}-\beta_1\frac{b}{a}\\ 0\end{cases}
$$

Since $\alpha$ is greater than one, $a<0$ and $b<0$. Thus the non-zero root is negative.

**L**emma 2: When $q \in (\frac{1}{1+\alpha}, \frac{1}{2})$, the gains to exploration over exploitation are decreasing and convex in $\mu$ for small $\mu$, increasing and concave for intermediate $\mu$, and decreasing and convex for large $\mu$. When $q \in \left(\frac{1}{2}, \frac{3\alpha+1}{4\alpha+1}\right)$, they are increasing and convex for small $\mu$ and decreasing and convex for large $\mu$. When $q > \frac{3\alpha+1}{4\alpha+1}$, they are increasing and concave for small $\mu$ and decreasing and convex for large $\mu$. When $q < \frac{1}{1+\alpha}$, they are decreasing and convex for small $\mu$ and increasing and concave for large $\mu$. In every case, the gains are zero when $\mu = 0$; when $q > \frac{1}{1+\alpha}$ ($q < \frac{1}{1+\alpha}$), they approach zero from above (below) as $\mu \to \infty$, holding $\beta_1$ fixed.

**P**roof:

Part (i): Low $\mu$. As $\mu \longrightarrow 0$:

$$\begin{aligned}
qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(\beta_1) &= q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1+\mu}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1+\mu}\right)-\left(\frac{2\beta_1}{2\beta_1+\mu}\right)\\
&\longrightarrow q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1}\right)-\left(\frac{2\beta_1}{2\beta_1}\right)\\
&= q+(1-q)-1=0
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\mu}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(\beta_1)\right] &= q\left(\frac{-(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^2}\right)+(1-q)\left(\frac{-\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^2}\right)+\frac{2\beta_1}{(2\beta_1+\mu)^2}\\
&\longrightarrow q\left(\frac{-1}{(1+\alpha)\beta_0}\right)+(1-q)\left(\frac{-1}{\left(1+\frac{1}{\alpha}\right)\beta_0}\right)+\frac{1}{2\beta_0}\\
&= \frac{-2q\left(1+\frac{1}{\alpha}\right)-2(1-q)(1+\alpha)+(1+\alpha)\left(1+\frac{1}{\alpha}\right)}{2(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_1}\\
&= \frac{-2\left(q+\frac{1}{\alpha}q\right)-2(1-q+\alpha-\alpha q)+\left(1+\alpha+\frac{1}{\alpha}+1\right)}{2(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_1}\\
&= \frac{2\alpha q-2\frac{1}{\alpha}q-\alpha+\frac{1}{\alpha}}{2(1+\alpha)\left(1+\frac{1}{\alpha}\right)\beta_1} = \frac{\left(\alpha^2-1\right)(2q-1)}{2(1+\alpha)^2\beta_1} \longrightarrow \begin{cases} 0^+ & \text{if } q>\frac{1}{2}\\ 0^- & \text{if } q<\frac{1}{2}\end{cases}
\end{aligned}$$

$$\frac{\partial^2}{\partial\mu^2}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(\beta_1)\right] = q\left(\frac{2(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^3}\right)+(1-q)\left(\frac{2\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^3}\right)-\frac{4\beta_1}{(2\beta_1+\mu)^3}$$

$$\text{num.} \longrightarrow \left(4q(1+\alpha)\left(1+\frac{1}{\alpha}\right)^3+4(1-q)\left(1+\frac{1}{\alpha}\right)(1+\alpha)^3-(1+\alpha)^3\left(1+\frac{1}{\alpha}\right)^3\right)4\beta_0^7 \begin{cases} >0 & \text{if } q<\frac{3\alpha+1}{4(\alpha+1)}\\ <0 & \text{if } q>\frac{3\alpha+1}{4(\alpha+1)}\end{cases}$$

Part (ii): High $\mu$. As $\mu \longrightarrow \infty$:

$$\begin{aligned}
qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(\beta_1) &= q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1+\mu}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1+\mu}\right)-\left(\frac{2\beta_1}{2\beta_1+\mu}\right)\\
&\longrightarrow \frac{1}{\mu}\left(q(1+\alpha)\beta_1+(1-q)\left(1+\frac{1}{\alpha}\right)\beta_1-2\beta_1\right)\\
&= \frac{1}{\mu}\beta_1\left(\alpha q-\frac{1}{\alpha}q-1+\frac{1}{\alpha}\right)\\
&= \frac{1}{\alpha\mu}\beta_1\left(\left(\alpha^2-1\right)q-(\alpha-1)\right)\\
&= \frac{\alpha-1}{\alpha\mu}\beta_1((1+\alpha)q-1)\longrightarrow \begin{cases}0^+ & \text{if } q>\frac{1}{1+\alpha}\\ 0^- & \text{if } q<\frac{1}{1+\alpha}\end{cases}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\mu}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(\beta_1)\right] &= q\left(\frac{-(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^2}\right)+(1-q)\left(\frac{-\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^2}\right)+\frac{2\beta_1}{(2\beta_1+\mu)^2}\\
&\longrightarrow \frac{1}{\mu^2}\left(-q(1+\alpha)\beta_1-(1-q)\left(1+\frac{1}{\alpha}\right)\beta_1+2\beta_1\right)\\
&= \frac{1}{\mu^2}\beta_1\left(-(q+\alpha q)-\left(1-q+\frac{1}{\alpha}-\frac{1}{\alpha}q\right)+2\right)\\
&= \frac{1}{\mu^2}\beta_1\left(-\alpha q+\frac{1}{\alpha}q+1-\frac{1}{\alpha}\right)\\
&= \frac{1}{\alpha\mu^2}\beta_1\left(-\left(\alpha^2-1\right)q+(\alpha-1)\right)\\
&= \frac{\alpha-1}{\alpha\mu^2}\beta_1(-(1+\alpha)q+1)\beta_1\longrightarrow \begin{cases}0^+ & \text{if } q<\frac{1}{1+\alpha}\\ 0^- & \text{if } q>\frac{1}{1+\alpha}\end{cases}
\end{aligned}$$

$$\frac{\partial^2}{\partial\mu^2}\left[qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F(\beta_1)\right] = q\left(\frac{2(1+\alpha)\beta_1}{((1+\alpha)\beta_1+\mu)^3}\right)+(1-q)\left(\frac{2\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)^3}\right)-\frac{4\beta_1}{(2\beta_1+\mu)^3},$$

$$\text{numerator} \longrightarrow \left(q\cdot(1+\alpha)+(1-q)\cdot\left(1+\tfrac{1}{\alpha}\right)-2\right)2\beta_1\mu^6 \begin{cases}>0 & \text{if } q>\frac{1}{1+\alpha}\\ <0 & \text{if } q<\frac{1}{1+\alpha}\end{cases}$$

Taken together, these asymptotics generate a curve with the shape described.

**P**roposition 2: When $q > \frac{1}{1+\alpha}$, there exists a unique level of competition $\mu_2^*$ at which the gains to exploration, relative to exploitation, are maximized.

**P**roof: Existence follows from lemma and continuity of the success function. Since the difference of the success function under exploration and exploitation is quadratic in $\mu$, it has at most two real roots, one of which is shown below to be zero, the other of which is shown to be positive if $q \in \left(\frac{1}{1+\alpha}, \frac{1}{2}\right)$ and negative otherwise. Given the shape described by the lemma, the value at which this difference is maximized must be unique.

To find the roots, set $qF\left(\beta_2^H\right)+(1-q)F\left(\beta_2^L\right)-F\left(\beta_1\right)=0$ and solve for $\mu$:

$$
\begin{aligned}
0 &= q\left(\frac{(1+\alpha)\beta_1}{(1+\alpha)\beta_1+\mu}\right)+(1-q)\left(\frac{\left(1+\frac{1}{\alpha}\right)\beta_1}{\left(1+\frac{1}{\alpha}\right)\beta_1+\mu}\right)-\left(\frac{2\beta_1}{2\beta_1+\mu}\right)\\
&= q(1+\alpha)\beta_1\left[\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)(2\beta_1+\mu)\right]\\
&\quad +(1-q)\left(1+\frac{1}{\alpha}\right)\beta_1\left[((1+\alpha)\beta_1+\mu)(2\beta_1+\mu)\right]\\
&\quad -2\beta_1\left[((1+\alpha)\beta_1+\mu)\left(\left(1+\frac{1}{\alpha}\right)\beta_1+\mu\right)\right]\\
&= \mu^2\beta_1\left[q(1+\alpha)+(1-q)\left(1+\frac{1}{\alpha}\right)-2\right]\\
&\quad +\mu\beta_1^2\left[q\left(2+\alpha+\frac{1}{\alpha}\right)+(1-q)\left(2+\alpha+\frac{1}{\alpha}\right)+2q(1+\alpha)+2(1-q)\left(1+\frac{1}{\alpha}\right)-2\left(2+\alpha+\frac{1}{\alpha}\right)\right]\\
&\quad +\beta_1^3\left[2q\left(2+\alpha+\frac{1}{\alpha}\right)+2(1-q)\left(2+\alpha+\frac{1}{\alpha}\right)-2\left(2+\alpha+\frac{1}{\alpha}\right)\right]\\
&= a\beta_1\mu^2+b\beta_1^2\mu+c\beta_1^3,
\end{aligned}
$$

where

$$
\begin{aligned}
a &= q+\alpha q+(1-q)+\frac{1}{\alpha}(1-q)-2=\alpha q+\frac{1}{\alpha}(1-q)-1=\frac{1}{\alpha}(\alpha-1)((1+\alpha)q-1)\begin{cases}>0 & \text{if } q>\frac{1}{1+\alpha}\\ <0 & \text{if } q<\frac{1}{1+\alpha}\end{cases}\\
b &= 2q+2\alpha q+2(1-q)+2\frac{1}{\alpha}(1-q)-\left(2+\alpha+\frac{1}{\alpha}\right)=\frac{1}{\alpha}\left(\alpha^2-1\right)(2q-1)\begin{cases}>0 & \text{if } q>\frac{1}{2}\\ <0 & \text{if } q<\frac{1}{2}\end{cases}\\
c &= 0
\end{aligned}
$$

By the quadratic formula, the roots are thus:

$$
\frac{-\left(b\beta_1^2\right)\pm\sqrt{\left(b\beta_1^2\right)^2-0}}{2(a\beta_1)}=\frac{-(b\beta_1)\pm-(b\beta_1)}{2a}=\begin{cases}-\beta_1\frac{b}{a}\\ 0\end{cases}
$$

When $q<\frac{1}{1+\alpha}$, $a<0$ and $b<0$, and the non-zero root is negative. When $q\in\left(\frac{1}{1+\alpha},\frac{1}{2}\right)$, $a>0$ and $b<0$, and the non-zero root is positive. When $q>\frac{1}{2}$, $a>0$ and $b>0$, and the non-zero root is negative.

**C**orollary: When $q < \frac{1}{1+\alpha}$, exploration will never be preferred to exploitation.

**P**roof: Follows from lemma, continuity of the success function, and results from the previous proof showing that when $q < \frac{1}{1+\alpha}$, there is no positive root for the difference of the success function for exploration and exploitation, such that this difference never becomes positive.

**P**roposition 3: At very low and very high $\mu$, the next-best alternative to exploration is abandonment. At intermediate $\mu$, the next-best alternative is exploitation.

**P**roof: Lemma 1 can be used to characterize the shape of the gains to exploration versus abandonment and exploitation versus abandonment, since in this model, exploitation is a special case of exploration, with $\alpha = 1$. The proof to Lemma 1 establishes that the gains to exploitation are zero when $\mu = 0$, increasing for small $\mu$, decreasing for large $\mu$, and approach zero from above as $\mu \to \infty$. Provided the prize-normalized cost of exploitation is not greater than the maximum of this function, the payoffs to exploitation will begin negative, turn positive, and finish negative, implying that abandonment (the IR constraint) is binding to exploration at low and high $\mu$ and exploitation (the IC constraint) is binding at intermediate $\mu$.

**P**roposition 4: When $q > \frac{1}{1+\alpha}$, there exists a unique level of competition $\mu^* \in [\mu_1^*, \mu_2^*]$ at which the gains to exploration are maximized relative to the player's next-best alternative.

**P**roof: Result follows from the first three propositions.

Figure I.B.1: Exploration, exploitation, and abandonment regions

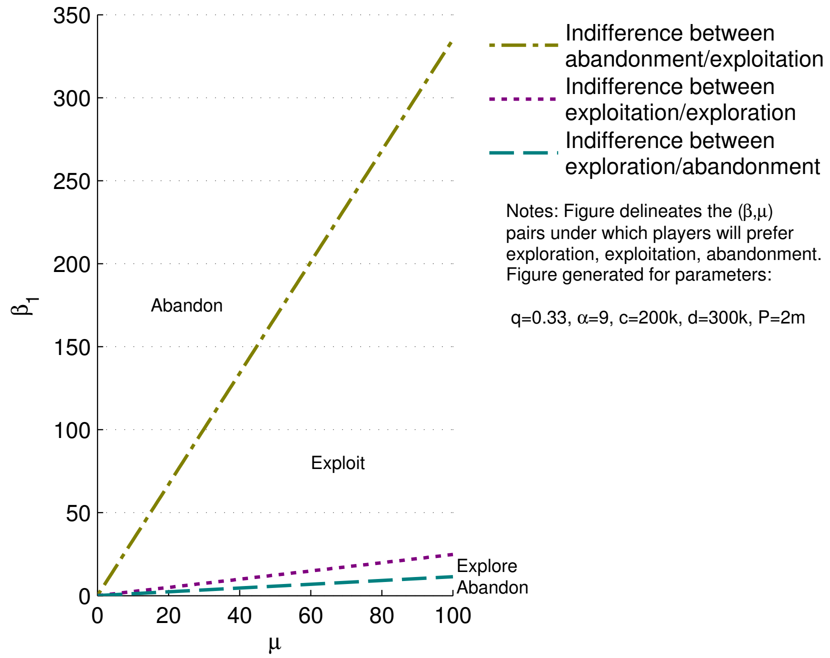
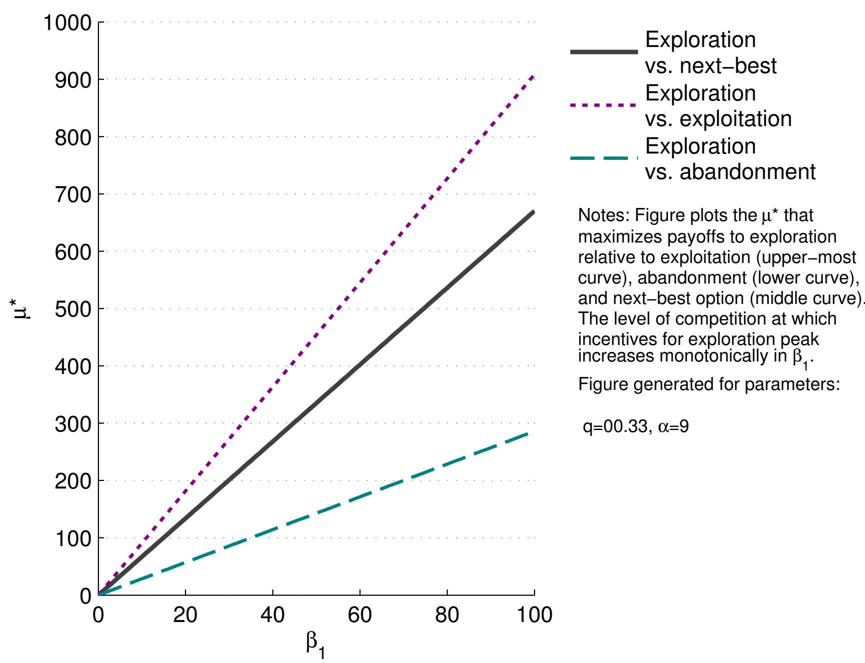



Figure I.B.2: $\mu^*$ at which benefits to exploration are maximized



## I.C 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.

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.

### 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 originality 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.

## 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.[1] 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[2] 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.

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 I.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.

[1] See http://www.hackerfactor.com/blog/archives/432-Looks-Like-It.html.
[2] See http://www.hackerfactor.com/blog/archives/529-Kind-of-Like-That.html.

## Why use algorithms?

There are three advantages to using algorithms over human judges. The first is that the algorithms provide a consistent, objective measure of similarity, whereas individuals can have significantly different, subjective perceptions of similarity in practice (Tirilly et al. 2012). This conclusion 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. The second advantage to algorithms over human judges is that algorithms can be directed to evaluate specific features of an image (in this case, the low frequency content or pixel intensity gradient), while human judges will see what they choose to see, and may be attuned to different features in different comparisons. The final advantage of algorithms is more obvious: they are cheap, taking only seconds to execute a comparison.

The evidence of disagreement in subjects' assessments of similarity nevertheless raises a deeper question: is it sensible to apply a uniform similarity measure in this setting? I argue that it is, for the following reasons. First, in both Tirilly et al. (2012) and the Mechanical Turk trials, respondents agreed on extremes, when images were either highly similar or highly dissimilar – in other words, it tends to be obvious when two images are near replicas, which is the margin of variation that matters most for this paper. Squire and Pun (1997) also found that *e*xpert subjects' assessments of similarity tend to agree at all levels; the designers in this paper could reasonably be classified as visual experts. Finally, divergence in opinion may result from the fact that subjects in the above studies were instructed to assess similarity as they perceive it, rather than in terms of specific features. If subjects were instructed to focus on specific features, they would likely tend to agree – not only with each other, but also with the computer.

**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." *P*roceedings 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." *P*roceedings of the Scandinavian Conference on Image Analysis, Lappeenranta, Finland.

## I.D Robustness Checks (1)

The following tables provide variants of the tables in Section I.3 estimating the effects of feedback and competition on experimentation, using the difference hash algorithm instead of the preferred, perceptual hash algorithm. These estimates serve as robustness checks to the principal empirical results of the paper, demonstrating that they are not sensitive to the procedure used to calculate similarity scores.

Table I.D.1 is a robustness check on Table I.7; Table I.D.2, on Table I.8; Table I.D.3, on Table I.9; Table I.D.4, on Table I.10; and Table I.D.5, on Table I.11. The results in these appendix tables are qualitatively and quantitatively similar to those in the body of the paper.

Table I.D.1: Similarity to player's previous designs (difference hash)

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Player's prior best rating==5 | 0.270*** | 0.253*** | 0.269*** | 0.256*** |
| | (0.086) | (0.087) | (0.087) | (0.086) |
| * 1+ competing 5-stars | -0.128** | -0.141** | -0.127** | -0.140** |
| | (0.058) | (0.056) | (0.058) | (0.056) |
| * prize value ($100s) | -0.038 | -0.047* | -0.034 | -0.046* |
| | (0.026) | (0.027) | (0.026) | (0.027) |
| Player's prior best rating==4 | 0.057*** | 0.025 | 0.064*** | 0.030 |
| | (0.020) | (0.021) | (0.020) | (0.021) |
| Player's prior best rating==3 | 0.027* | 0.011 | 0.035** | 0.016 |
| | (0.017) | (0.017) | (0.017) | (0.017) |
| Player's prior best rating==2 | -0.004 | -0.012 | 0.003 | -0.008 |
| | (0.020) | (0.020) | (0.020) | (0.020) |
| One or more competing 5-stars | -0.011 | -0.022 | -0.011 | -0.022 |
| | (0.022) | (0.024) | (0.022) | (0.023) |
| Days remaining | -0.004 | 0.001 | -0.004* | 0.001 |
| | (0.003) | (0.007) | (0.003) | (0.007) |
| Constant | 0.508*** | 0.454*** | 0.512*** | 0.461*** |
| | (0.139) | (0.160) | (0.139) | (0.159) |
| N | 5075 | 5075 | 5075 | 5075 |
| $R^2$ | 0.48 | 0.48 | 0.48 | 0.48 |
| Controls | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes |
| Forthcoming ratings | No | No | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of a design's maximum 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. The mean value of this variable in the sample is 0.58 (s.d. 0.28). Columns (2) and (4) control for time of submission and number of previous designs entered by the player and her competitors. Columns (3) and (4) additionally control for the best $f$orthcoming rating on the player's not-yet-rated designs. 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. SEs clustered by player in parentheses.

Table I.D.2: Similarity to player's best previously-rated designs & intra-batch similarity (d. hash)

| | Designs | | Batches (uwtd.) | | Batches (wtd.) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Player's prior best==5 | 0.244* | 0.242* | 0.224 | 0.246 | 0.236 | 0.260 |
| | (0.131) | (0.141) | (0.299) | (0.293) | (0.286) | (0.281) |
| * 1+ competing 5-stars | -0.168* | -0.177** | -0.327** | -0.324** | -0.314** | -0.308** |
| | (0.086) | (0.087) | (0.146) | (0.144) | (0.147) | (0.145) |
| * prize value ($100s) | -0.018 | -0.024 | -0.022 | -0.023 | -0.025 | -0.026 |
| | (0.038) | (0.042) | (0.093) | (0.092) | (0.087) | (0.085) |
| Player's prior best==4 | 0.066* | 0.049 | -0.016 | -0.003 | -0.012 | 0.004 |
| | (0.039) | (0.041) | (0.031) | (0.032) | (0.029) | (0.031) |
| Player's prior best==3 | 0.044 | 0.033 | 0.011 | 0.019 | 0.010 | 0.020 |
| | (0.038) | (0.039) | (0.033) | (0.035) | (0.031) | (0.032) |
| Player's prior best==2 | 0.014 | 0.007 | -0.019 | -0.012 | -0.021 | -0.014 |
| | (0.040) | (0.040) | (0.047) | (0.049) | (0.045) | (0.045) |
| 1+ competing 5-stars | -0.012 | -0.019 | -0.018 | -0.018 | -0.014 | -0.017 |
| | (0.031) | (0.032) | (0.033) | (0.034) | (0.031) | (0.033) |
| Days remaining | 0.002 | 0.001 | -0.000 | -0.002 | -0.000 | 0.001 |
| | (0.003) | (0.009) | (0.004) | (0.008) | (0.004) | (0.009) |
| Constant | 0.844*** | 0.863*** | 0.646*** | 0.673*** | 0.672*** | 0.661*** |
| | (0.152) | (0.173) | (0.121) | (0.156) | (0.096) | (0.128) |
| N | 3871 | 3871 | 1987 | 1987 | 1987 | 1987 |
| $R^2$ | 0.53 | 0.53 | 0.59 | 0.59 | 0.59 | 0.59 |
| Controls | No | Yes | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations in Columns (1) and (2) are designs, and 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). Observations in Columns (3) to (6) 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), and 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 0.72 (s.d. 0.27). Columns (5) and (6) weight the batch regressions by batch size. All columns control for the time of submission 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. SEs clustered by player in parentheses.

Table I.D.3: Change in similarity to player's best previously-rated designs (d. hash)

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Δ(Player's best==5) | 0.657*** | 0.680*** | 0.687** | 0.659*** | 0.693*** | 0.693*** |
| | (0.218) | (0.256) | (0.269) | (0.217) | (0.256) | (0.267) |
| * 1+ competing 5-stars | -0.347** | -0.374* | -0.362* | -0.350** | -0.379* | -0.368* |
| | (0.174) | (0.206) | (0.218) | (0.174) | (0.206) | (0.218) |
| * prize value ($100s) | -0.049 | -0.060 | -0.063 | -0.048 | -0.063 | -0.064 |
| | (0.046) | (0.055) | (0.057) | (0.046) | (0.055) | (0.057) |
| Δ(Player's best==4) | 0.262*** | 0.236*** | 0.231*** | 0.262*** | 0.237*** | 0.232*** |
| | (0.070) | (0.081) | (0.086) | (0.070) | (0.081) | (0.086) |
| Δ(Player's best==3) | 0.192*** | 0.169** | 0.161** | 0.192*** | 0.169** | 0.162** |
| | (0.062) | (0.073) | (0.077) | (0.063) | (0.073) | (0.077) |
| Δ(Player's best==2) | 0.132** | 0.110 | 0.104 | 0.131** | 0.110 | 0.104 |
| | (0.058) | (0.067) | (0.071) | (0.058) | (0.067) | (0.071) |
| 1+ competing 5-stars | -0.005 | -0.000 | -0.005 | 0.001 | 0.000 | -0.001 |
| | (0.016) | (0.016) | (0.025) | (0.018) | (0.016) | (0.029) |
| Days remaining | -0.000 | -0.000 | 0.000 | -0.007 | -0.005 | -0.009 |
| | (0.002) | (0.002) | (0.003) | (0.005) | (0.004) | (0.009) |
| Constant | -0.012 | -0.013 | -0.237*** | 0.058 | 0.038 | -0.169** |
| | (0.010) | (0.010) | (0.045) | (0.050) | (0.047) | (0.072) |
| N | 2694 | 2694 | 2694 | 2694 | 2694 | 2694 |
| $R^2$ | 0.04 | 0.10 | 0.13 | 0.04 | 0.10 | 0.13 |
| Controls | No | No | No | Yes | Yes | Yes |
| Contest FEs | Yes | No | Yes | Yes | No | Yes |
| Player FEs | No | Yes | Yes | No | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of the *c*hange 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 experimenting to copying (and a value of -1, the converse). The mean value of this variable in the sample is -0.01 (s.d. 0.25). Columns (4) to (6) control for time of submission and number of previous designs entered by the player and 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. SEs clustered by player in parentheses.

Table I.D.4: Similarity to player's best not-yet-rated designs (placebo test; using d. hash)

| | Similarity to forthcoming | | | Residual |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Player's best forthcoming rating==5 | 0.512 | 0.215 | 0.122 | 0.238 |
| | (0.447) | (0.214) | (0.202) | (0.207) |
| * 1+ competing 5-stars | -0.220 | -0.111 | -0.052 | -0.108 |
| | (0.262) | (0.144) | (0.146) | (0.145) |
| * prize value ($100s) | -0.122 | -0.026 | -0.006 | -0.024 |
| | (0.089) | (0.049) | (0.046) | (0.051) |
| Player's best forthcoming rating==4 | 0.105 | 0.105 | 0.109 | 0.122 |
| | (0.085) | (0.113) | (0.117) | (0.117) |
| Player's best forthcoming rating==3 | 0.093 | 0.103 | 0.101 | 0.121 |
| | (0.057) | (0.135) | (0.140) | (0.143) |
| Player's best forthcoming rating==2 | 0.045 | 0.049 | 0.048 | 0.077 |
| | (0.050) | (0.136) | (0.142) | (0.139) |
| One or more competing 5-stars | -0.076 | -0.079 | -0.081 | -0.075 |
| | (0.081) | (0.122) | (0.126) | (0.131) |
| Days remaining | -0.019 | 0.003 | 0.006 | 0.004 |
| | (0.030) | (0.055) | (0.056) | (0.062) |
| Constant | 0.998*** | 0.452 | 0.348 | -0.179 |
| | (0.170) | (0.473) | (0.513) | (0.489) |
| N | 1147 | 577 | 577 | 577 |
| $R^2$ | 0.69 | 0.87 | 0.88 | 0.69 |
| Controls | Yes | Yes | Yes | Yes |
| Contest FEs | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes |

Notes: Table provides a test of the effects of not-yet-available feedback on players' experimentation. Observations are designs. Dependent variable in Columns (1) to (3) is a continuous measure of a design's similarity to the best designs that the player has previously entered and has yet to *b*ut 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 in the sample is 0.50 (s.d. 0.29). If players depend on sponsors' ratings for signals of quality, then forthcoming ratings should have no effect on current experimentation. The results of Column (1) suggest this may not be the case; however, similarity to an unrated design may actually be the result of both these designs being tweaks on a third design. To account for this possibility, Column (2) controls for the given design's similarity to the best previously-rated design, the best not-yet-rated design's similarity to the best previously-rated design, and their interaction. Column (3) allows these controls to vary by the best 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 placebo that is not explained by jointly-occurring imitation of a third design. All columns control for time of submission 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. SEs clustered by player in parentheses.

Table I.D.5: Similarity/change in similarity to competitors' best previously-rated designs (d. hash)

| | (1) | (2) | | (3) | (4) |
|---|---|---|---|---|---|
| Competing best==5 | -0.168*** | -0.127*** | Δ(Competing best==5) | -0.064 | -0.062 |
| | -0.033 | -0.034 | | -0.096 | (0.096) |
| * 1+ own 5-stars | -0.018 | -0.017 | * 1+ own 5-stars | 0.006 | 0.005 |
| | -0.017 | -0.017 | | -0.066 | (0.066) |
| * prize value ($100s) | 0.025*** | 0.005 | * prize value ($100s) | 0.001 | -0.000 |
| | -0.006 | -0.006 | | -0.018 | (0.018) |
| Competing best==4 | 0.027 | 0.033 | Δ(Competing best==4) | 0.069 | 0.069 |
| | -0.025 | -0.025 | | -0.076 | (0.076) |
| Competing best==3 | 0.032 | 0.045* | Δ(Competing best==3) | 0.068 | 0.069 |
| | -0.025 | -0.025 | | -0.075 | (0.076) |
| Competing best==2 | -0.052* | -0.052* | Δ(Competing best==2) | 0.014 | 0.016 |
| | -0.03 | -0.029 | | -0.077 | (0.077) |
| One or more own 5-stars | 0.004 | 0.005 | One or more own 5-stars | 0.016 | 0.020 |
| | -0.027 | -0.03 | | -0.012 | (0.013) |
| Days remaining | -0.011*** | 0.008*** | Days remaining | 0.001 | 0.004 |
| | -0.001 | -0.002 | | -0.001 | (0.003) |
| Constant | 0.537*** | 0.396*** | Constant | 0.035 | 0.016 |
| | -0.094 | -0.097 | | -0.114 | (0.118) |
| N | 9709 | 9709 | N | 6065 | 6065 |
| $R^2$ | 0.54 | 0.54 | $R^2$ | 0.14 | 0.15 |
| Controls | No | Yes | Controls | No | Yes |
| Contest FEs | Yes | Yes | Contest FEs | Yes | Yes |
| Player FEs | Yes | Yes | Player FEs | Yes | Yes |

Notes: Table provides a test of players' ability to discern the quality of, and then imitate, competing designs. Observations are designs. Dependent variable in Columns (1) and (2) is a continuous measure of the design's similarity to the highest-rated preceding entries by *o*ther players, taking values in [0,1], where a value of 1 indicates that the design is identical to another. The mean value in the sample is 0.33 (s.d. 0.21). Dependent variable in Columns (3) and (4) is a continuous measure of the *c*hange in designs' similarity to the highest- rated preceding entries by other players, taking values in [-1,1], where a value of 0 indicates that the player's current design is equally similar to the best competing design as was her previous design, and a value of 1 indicates that the player transitioned fully from experimenting to copying (and a value of -1, the converse). The mean value of this variable in the sample is 0.00 (s.d. 0.15). In general, players are provided only the distribution of ratings on competing designs; ratings of specific competing designs are not observed. Results in this table test whether players can nevertheless identify and imitate leading competition. Columns (2) and (4) control for time of submission 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. Robust SEs in parentheses.

## I.E Robustness Checks (2)

In additional robustness checks, I show that competition has a constant effect on high-performing players' tendency to experiment. Tables I.E.1 to I.E.3 demonstrate this result with the perceptual hash similarity measures, and Tables I.E.4 to I.E.6 do so with the difference hash measures. In all cases, I estimate differential effects for one vs. multiple top-rated, competing designs and find no differential effect.

Table I.E.1: Similarity to player's previous designs (p. hash)

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Player's prior best rating==5 | 0.269*** | 0.250*** | 0.268*** | 0.253*** |
| | (0.087) | (0.087) | (0.088) | (0.086) |
| * 1+ competing 5-stars | -0.102 | -0.110 | -0.101 | -0.108 |
| | (0.087) | (0.080) | (0.087) | (0.081) |
| * 2+ competing 5-stars | -0.034 | -0.039 | -0.034 | -0.041 |
| | (0.094) | (0.080) | (0.093) | (0.081) |
| * prize value ($100s) | -0.037 | -0.046* | -0.034 | -0.045* |
| | (0.026) | (0.027) | (0.027) | (0.027) |
| Player's prior best rating==4 | 0.058*** | 0.027 | 0.066*** | 0.032 |
| | (0.020) | (0.021) | (0.020) | (0.021) |
| Player's prior best rating==3 | 0.028* | 0.013 | 0.036** | 0.018 |
| | (0.017) | (0.017) | (0.017) | (0.017) |
| Player's prior best rating==2 | -0.004 | -0.012 | 0.003 | -0.008 |
| | (0.020) | (0.020) | (0.020) | (0.020) |
| One or more competing 5-stars | -0.042 | -0.044 | -0.043 | -0.045 |
| | (0.036) | (0.038) | (0.036) | (0.037) |
| Two or more competing 5-stars | 0.049 | 0.036 | 0.050 | 0.037 |
| | (0.037) | (0.040) | (0.038) | (0.040) |
| Days remaining | -0.004 | 0.001 | -0.004 | 0.001 |
| | (0.003) | (0.007) | (0.002) | (0.007) |
| Constant | 0.506*** | 0.460*** | 0.510*** | 0.467*** |
| | (0.139) | (0.160) | (0.138) | (0.159) |
| N | 5075 | 5075 | 5075 | 5075 |
| $R^2$ | 0.48 | 0.48 | 0.48 | 0.48 |
| Controls | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes |
| Forthcoming ratings | No | No | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of a design's maximum 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. The mean value of this variable in the sample is 0.32 (s.d. 0.27). Columns (2) and (4) control for time of submission and number of previous designs entered by the player and her competitors. Columns (3) and (4) additionally control for the best $f$orthcoming rating on the player's not-yet-rated designs. 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. SEs clustered by player.

Table I.E.2: Similarity to player's best previously-rated designs & intra-batch similarity (p. hash)

| | Designs | | Batches (uwtd.) | | Batches (wtd.) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Player's prior best==5 | 0.239* | 0.234* | 0.239 | 0.271 | 0.253 | 0.285 |
| | (0.132) | (0.141) | (0.317) | (0.312) | (0.300) | (0.295) |
| * 1+ competing 5-stars | -0.144 | -0.149 | -0.327** | -0.328** | -0.328** | -0.327** |
| | (0.140) | (0.138) | (0.143) | (0.142) | (0.131) | (0.131) |
| * 2+ competing 5-stars | -0.032 | -0.036 | 0.006 | 0.009 | 0.025 | 0.030 |
| | (0.143) | (0.134) | (0.189) | (0.186) | (0.173) | (0.172) |
| * prize value ($100s) | -0.016 | -0.022 | -0.026 | -0.031 | -0.030 | -0.033 |
| | (0.038) | (0.042) | (0.100) | (0.099) | (0.092) | (0.090) |
| Player's prior best==4 | 0.067* | 0.050 | -0.015 | -0.002 | -0.012 | 0.004 |
| | (0.039) | (0.041) | (0.031) | (0.032) | (0.029) | (0.031) |
| Player's prior best==3 | 0.044 | 0.033 | 0.012 | 0.019 | 0.011 | 0.020 |
| | (0.038) | (0.039) | (0.033) | (0.035) | (0.031) | (0.032) |
| Player's prior best==2 | 0.013 | 0.006 | -0.018 | -0.011 | -0.020 | -0.012 |
| | (0.040) | (0.040) | (0.047) | (0.048) | (0.044) | (0.045) |
| 1+ competing 5-stars | -0.050 | -0.050 | 0.037 | 0.038 | 0.031 | 0.030 |
| | (0.045) | (0.045) | (0.041) | (0.042) | (0.037) | (0.038) |
| 2+ competing 5-stars | 0.058 | 0.052 | -0.089* | -0.096* | -0.076 | -0.082 |
| | (0.049) | (0.051) | (0.052) | (0.055) | (0.048) | (0.051) |
| Days remaining | 0.002 | 0.001 | -0.001 | -0.003 | -0.001 | -0.000 |
| | (0.003) | (0.009) | (0.004) | (0.009) | (0.004) | (0.009) |
| Constant | 0.841*** | 0.870*** | 0.649*** | 0.677*** | 0.675*** | 0.666*** |
| | (0.152) | (0.174) | (0.120) | (0.156) | (0.095) | (0.128) |
| N | 3871 | 3871 | 1987 | 1987 | 1987 | 1987 |
| $R^2$ | 0.53 | 0.54 | 0.59 | 0.59 | 0.59 | 0.59 |
| Controls | No | Yes | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations in Columns (1) and (2) are designs, and 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). Observations in Columns (3) to (6) 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), and 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 0.48 (s.d. 0.32). Columns (5) and (6) weight the batch regressions by batch size. All columns control for the time of submission 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. SEs clustered by player in parentheses.

Table I.E.3: Change in similarity to player's best previously-rated designs (p. hash)

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| $\Delta$(Player's best==5) | 0.656*** | 0.685*** | 0.693*** | 0.657*** | 0.695*** | 0.695*** |
| | (0.218) | (0.253) | (0.265) | (0.217) | (0.253) | (0.264) |
| * 1+ competing 5-stars | -0.398* | -0.449* | -0.425* | -0.400* | -0.456* | -0.429* |
| | (0.210) | (0.235) | (0.255) | (0.211) | (0.236) | (0.256) |
| * 2+ competing 5-stars | 0.070 | 0.106 | 0.090 | 0.068 | 0.109 | 0.088 |
| | (0.154) | (0.167) | (0.185) | (0.155) | (0.167) | (0.185) |
| * prize value ($100s) | -0.048 | -0.060 | -0.064 | -0.047 | -0.062 | -0.063 |
| | (0.046) | (0.053) | (0.055) | (0.045) | (0.053) | (0.055) |
| $\Delta$(Player's best==4) | 0.264*** | 0.240*** | 0.235*** | 0.264*** | 0.241*** | 0.237*** |
| | (0.070) | (0.081) | (0.086) | (0.070) | (0.081) | (0.085) |
| $\Delta$(Player's best==3) | 0.194*** | 0.174** | 0.165** | 0.194*** | 0.173** | 0.166** |
| | (0.063) | (0.073) | (0.077) | (0.063) | (0.074) | (0.077) |
| $\Delta$(Player's best==2) | 0.134** | 0.114* | 0.108 | 0.133** | 0.114* | 0.108 |
| | (0.058) | (0.068) | (0.071) | (0.058) | (0.068) | (0.071) |
| 1+ competing 5-stars | -0.013 | -0.020 | -0.031 | -0.008 | -0.018 | -0.026 |
| | (0.029) | (0.035) | (0.047) | (0.029) | (0.036) | (0.048) |
| 2+ competing 5-stars | 0.011 | 0.028 | 0.041 | 0.015 | 0.026 | 0.046 |
| | (0.027) | (0.037) | (0.047) | (0.028) | (0.038) | (0.048) |
| Days remaining | -0.000 | -0.000 | 0.001 | -0.007 | -0.005 | -0.009 |
| | (0.002) | (0.002) | (0.003) | (0.005) | (0.004) | (0.009) |
| Constant | -0.013 | -0.014 | -0.243*** | 0.059 | 0.040 | -0.171** |
| | (0.010) | (0.010) | (0.044) | (0.049) | (0.046) | (0.073) |
| N | 2694 | 2694 | 2694 | 2694 | 2694 | 2694 |
| $R^2$ | 0.04 | 0.10 | 0.13 | 0.04 | 0.10 | 0.13 |
| Controls | No | No | No | Yes | Yes | Yes |
| Contest FEs | Yes | No | Yes | Yes | No | Yes |
| Player FEs | No | Yes | Yes | No | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of the *c*hange 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 experimenting to copying (and a value of -1, the converse). The mean value of this variable in the sample is -0.00 (s.d. 0.23). Columns (4) to (6) control for time of submission and number of previous designs entered by the player and 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. SEs clustered by player in parentheses.

Table I.E.4: Similarity to player's previous designs (d. hash)

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Player's prior best rating==5 | 0.269*** | 0.250*** | 0.268*** | 0.253*** |
| | (0.087) | (0.087) | (0.088) | (0.086) |
| * 1+ competing 5-stars | -0.102 | -0.110 | -0.101 | -0.108 |
| | (0.087) | (0.080) | (0.087) | (0.081) |
| * 2+ competing 5-stars | -0.034 | -0.039 | -0.034 | -0.041 |
| | (0.094) | (0.080) | (0.093) | (0.081) |
| * prize value ($100s) | -0.037 | -0.046* | -0.034 | -0.045* |
| | (0.026) | (0.027) | (0.027) | (0.027) |
| Player's prior best rating==4 | 0.058*** | 0.027 | 0.066*** | 0.032 |
| | (0.020) | (0.021) | (0.020) | (0.021) |
| Player's prior best rating==3 | 0.028* | 0.013 | 0.036** | 0.018 |
| | (0.017) | (0.017) | (0.017) | (0.017) |
| Player's prior best rating==2 | -0.004 | -0.012 | 0.003 | -0.008 |
| | (0.020) | (0.020) | (0.020) | (0.020) |
| One or more competing 5-stars | -0.042 | -0.044 | -0.043 | -0.045 |
| | (0.036) | (0.038) | (0.036) | (0.037) |
| Two or more competing 5-stars | 0.049 | 0.036 | 0.050 | 0.037 |
| | (0.037) | (0.040) | (0.038) | (0.040) |
| Days remaining | -0.004 | 0.001 | -0.004 | 0.001 |
| | (0.003) | (0.007) | (0.002) | (0.007) |
| Constant | 0.506*** | 0.460*** | 0.510*** | 0.467*** |
| | (0.139) | (0.160) | (0.138) | (0.159) |
| N | 5075 | 5075 | 5075 | 5075 |
| $R^2$ | 0.48 | 0.48 | 0.48 | 0.48 |
| Controls | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes |
| Forthcoming ratings | No | No | Yes | Yes |

Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of a design's maximum 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. The mean value of this variable in the sample is 0.58 (s.d. 0.28). Columns (2) and (4) control for time of submission and number of previous designs entered by the player and her competitors. Columns (3) and (4) additionally control for the best $f$orthcoming rating on the player's not-yet-rated designs. 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. SEs clustered by player.

Table I.E.5: Similarity to player's best previously-rated designs & intra-batch similarity (d. hash)

| | Designs | | Batches (uwtd.) | | Batches (wtd.) | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Player's prior best==5 | 0.239* | 0.234* | 0.239 | 0.271 | 0.253 | 0.285 |
| | (0.132) | (0.141) | (0.317) | (0.312) | (0.300) | (0.295) |
| * 1+ competing 5-stars | -0.144 | -0.149 | -0.327** | -0.328** | -0.328** | -0.327** |
| | (0.140) | (0.138) | (0.143) | (0.142) | (0.131) | (0.131) |
| * 2+ competing 5-stars | -0.032 | -0.036 | 0.006 | 0.009 | 0.025 | 0.030 |
| | (0.143) | (0.134) | (0.189) | (0.186) | (0.173) | (0.172) |
| * prize value ($100s) | -0.016 | -0.022 | -0.026 | -0.031 | -0.030 | -0.033 |
| | (0.038) | (0.042) | (0.100) | (0.099) | (0.092) | (0.090) |
| Player's prior best==4 | 0.067* | 0.050 | -0.015 | -0.002 | -0.012 | 0.004 |
| | (0.039) | (0.041) | (0.031) | (0.032) | (0.029) | (0.031) |
| Player's prior best==3 | 0.044 | 0.033 | 0.012 | 0.019 | 0.011 | 0.020 |
| | (0.038) | (0.039) | (0.033) | (0.035) | (0.031) | (0.032) |
| Player's prior best==2 | 0.013 | 0.006 | -0.018 | -0.011 | -0.020 | -0.012 |
| | (0.040) | (0.040) | (0.047) | (0.048) | (0.044) | (0.045) |
| 1+ competing 5-stars | -0.050 | -0.050 | 0.037 | 0.038 | 0.031 | 0.030 |
| | (0.045) | (0.045) | (0.041) | (0.042) | (0.037) | (0.038) |
| 2+ competing 5-stars | 0.058 | 0.052 | -0.089* | -0.096* | -0.076 | -0.082 |
| | (0.049) | (0.051) | (0.052) | (0.055) | (0.048) | (0.051) |
| Days remaining | 0.002 | 0.001 | -0.001 | -0.003 | -0.001 | -0.000 |
| | (0.003) | (0.009) | (0.004) | (0.009) | (0.004) | (0.009) |
| Constant | 0.841*** | 0.870*** | 0.649*** | 0.677*** | 0.675*** | 0.666*** |
| | (0.152) | (0.174) | (0.120) | (0.156) | (0.095) | (0.128) |
| N | 3871 | 3871 | 1987 | 1987 | 1987 | 1987 |
| $R^2$ | 0.53 | 0.54 | 0.59 | 0.59 | 0.59 | 0.59 |
| Controls | No | Yes | No | Yes | No | Yes |
| Contest FEs | Yes | Yes | Yes | Yes | Yes | Yes |
| Player FEs | Yes | Yes | Yes | Yes | Yes | Yes |

Notes: Table shows the effects of feedback on players'experimentation. Observations in Columns (1) and (2) are designs, and 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). Observations in Columns (3) to (6) 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), and 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 0.72 (s.d. 0.27). Columns (5) and (6) weight the batch regressions by batch size. All columns control for the time of submission 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. SEs clustered by player in parentheses.

Table I.E.6: Change in similarity to player's best previously-rated designs (d. hash)

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Δ(Player's best==5) | 0.656*** | 0.685*** | 0.693*** | 0.657*** | 0.695*** | 0.695*** |
| | (0.218) | (0.253) | (0.265) | (0.217) | (0.253) | (0.264) |
| * 1+ competing 5-stars | -0.398* | -0.449* | -0.425* | -0.400* | -0.456* | -0.429* |
| | (0.210) | (0.235) | (0.255) | (0.211) | (0.236) | (0.256) |
| * 2+ competing 5-stars | 0.070 | 0.106 | 0.090 | 0.068 | 0.109 | 0.088 |
| | (0.154) | (0.167) | (0.185) | (0.155) | (0.167) | (0.185) |
| * prize value ($100s) | -0.048 | -0.060 | -0.064 | -0.047 | -0.062 | -0.063 |
| | (0.046) | (0.053) | (0.055) | (0.045) | (0.053) | (0.055) |
| Δ(Player's best==4) | 0.264*** | 0.240*** | 0.235*** | 0.264*** | 0.241*** | 0.237*** |
| | (0.070) | (0.081) | (0.086) | (0.070) | (0.081) | (0.085) |
| Δ(Player's best==3) | 0.194*** | 0.174** | 0.165** | 0.194*** | 0.173** | 0.166** |
| | (0.063) | (0.073) | (0.077) | (0.063) | (0.074) | (0.077) |
| Δ(Player's best==2) | 0.134** | 0.114* | 0.108 | 0.133** | 0.114* | 0.108 |
| | (0.058) | (0.068) | (0.071) | (0.058) | (0.068) | (0.071) |
| 1+ competing 5-stars | -0.013 | -0.020 | -0.031 | -0.008 | -0.018 | -0.026 |
| | (0.029) | (0.035) | (0.047) | (0.029) | (0.036) | (0.048) |
| 2+ competing 5-stars | 0.011 | 0.028 | 0.041 | 0.015 | 0.026 | 0.046 |
| | (0.027) | (0.037) | (0.047) | (0.028) | (0.038) | (0.048) |
| Days remaining | -0.000 | -0.000 | 0.001 | -0.007 | -0.005 | -0.009 |
| | (0.002) | (0.002) | (0.003) | (0.005) | (0.004) | (0.009) |
| Constant | -0.013 | -0.014 | -0.243*** | 0.059 | 0.040 | -0.171** |
| | (0.010) | (0.010) | (0.044) | (0.049) | (0.046) | (0.073) |
| N | 2694 | 2694 | 2694 | 2694 | 2694 | 2694 |
| $R^2$ | 0.04 | 0.10 | 0.13 | 0.04 | 0.10 | 0.13 |
| Controls | No | No | No | Yes | Yes | Yes |
| Contest FEs | Yes | No | Yes | Yes | No | Yes |
| Player FEs | No | Yes | Yes | No | Yes | Yes |

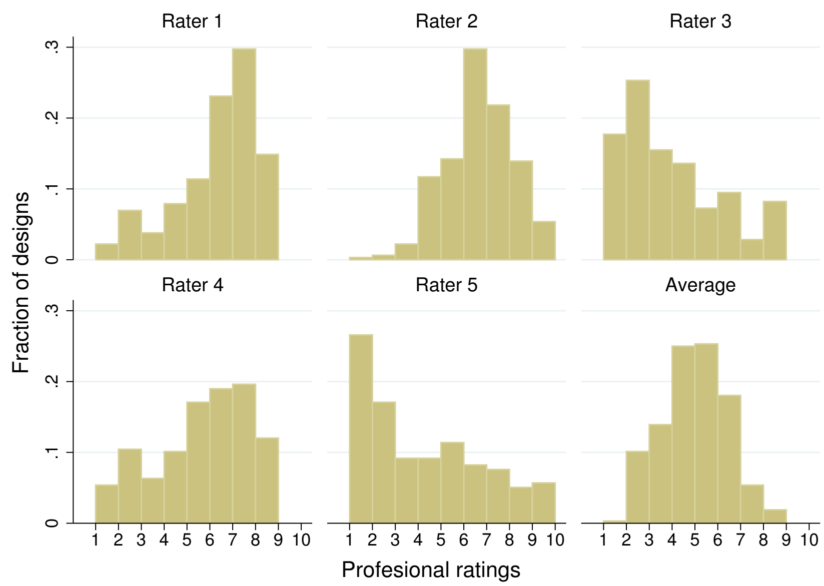
Notes: Table shows the effects of feedback on players' experimentation. Observations are designs. Dependent variable is a continuous measure of the *c*hange in designs' similarity to the highest-rated preceding entry by the same player, taking values in [-1,1], where a value of 0 indicates that the player's 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 experimenting to copying (and a value of -1, the converse). The mean value of this variable in the sample is -0.01 (s.d. 0.25). Columns (4) to (6) control for time of submission and number of previous designs entered by the player and 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. SEs clustered by player in parentheses.

## I.F Collection of Professional Ratings

The panelists that participated in the ratings exercise were recruited through the author's personal and professional networks and hired at their freelance rates. All have formal training and experience in commercial 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 created and managed on Amazon Mechanical Turk. 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 (as they appear 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 I.F.1 provides the distribution of ratings from each of the five panelists and their average.

Figure I.F.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 what the panelist intended with these ratings, why such a high proportion was given the lowest rating, and 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 results of the paper are nevertheless robust to including ratings from this panelist that lie above the lower bound.

## I.G Discussion of Social Welfare

Continued experimentation is always in the sponsor's best interest. But the implications for players' welfare and social welfare are ambiguous: the social benefits to innovation can exceed the private benefits, and the social costs will always be greater than the individual designer's cost, due to the negative externalities from competitive business stealing. In this appendix, I elaborate on the welfare implications of prize competition as a mechanism for procuring innovation, focusing on the model of Section I.1.

Whether a player's effort is socially optimal depends on the incremental value it generates and the cost of the effort incurred. By this criterion, even tweaks can be desirable, since they come with a new draw of the stochastic component ($\varepsilon$) of the innovation's value. To formalize the argument, let $V_{jt}$ be the value of the most valuable design to-date prior to the $t$-th design by player $j$, and let $\nu_{jt} = \ln(\beta_{jt}) + \varepsilon_{jt}$ continue to denote the value of design $jt$, as in the body of the paper. A new design will only be socially optimal if it is higher-value than $V_{jt}$, which occurs with probability $Pr(\nu_{jt} > V_{jt})$; otherwise, it will be discarded. Letting $\Pi^S$ denote social welfare, we can write the expected welfare gains as follows:

$$\begin{aligned} E\left[\Delta\Pi^S\right] &= \underbrace{E\left[\nu_{jt} - V_{jt} \mid \nu_{jt} > V_{jt}\right] \cdot Pr\left(\nu_{jt} > V_{jt}\right)}_{\text{Expected incremental value of an upgrade}} + \underbrace{0 \cdot Pr\left(\nu_{jt} \leq V_{jt}\right)}_{\text{Design discarded}} - \text{cost of effort} \\ &= E\left[\ln(\beta_{jt}) + \varepsilon_{jt} - V_{jt} \mid \varepsilon_{jt} > V_{jt} - \ln(\beta_{jt})\right] \cdot Pr\left(\varepsilon_{jt} > V_{jt} - \ln(\beta_{jt})\right) - \text{cost} \ , \end{aligned}$$

which may be greater than or less than zero. Note that this expression omits the change in each players' expected earnings, which offset each other – the net effect is strictly a function of the beneficiary's gains and the player's private costs. The condition for a socially optimal decision-rule thus reduces to whether the innovation value exceeds the private cost of innovating, be it radical or incremental.

Private choices can deviate from the social optimum under a multitude of circumstances. Because the private benefits are bounded at the dollar prize, whereas the social benefits are unbounded – and potentially quite large, if the fruits of innovation are enjoyed by an entire society – innovation can be inefficiently low unless the prize fully reflects the social value of the innovation. This is more likely to occur in explicit tournaments than in market settings, where the prize is monopoly rents, the size of which are dynamically determined by the level and shape of demand. On the other hand, rent-seeking motives may encourage players to exert effort that increases their expected earnings but yields no net value. A more precise understanding of social welfare would require a specific empirical example or parametrization.

## I.H Summary of Companion Paper: Gross (2015b)

In a companion paper, I use a sample of 4,294 logo design contests from the same setting to study the effects of feedback on tournament outcomes. Whereas the present paper studies the effects of feedback on the creative process, in the companion paper I focus on the effects of feedback on the quality of creative outcomes; the distinction is that of the process of innovation versus its result, which in this case is a copyrightable product. Gross (2015b) shows that feedback affects design quality via two channels: a selection effect, whereby unsuccessful players are driven to exit and successful players continue, and a direction effect, which guides continuing players towards better designs. In the paper, I use a combination of structural estimation and counterfactual simulations to establish that improvements in quality resulting from feedback accrue entirely as a result of direction rather than selection. These findings imply that successful innovation in this setting requires continuous learning and improvement substantially more than talent or luck.

To highlight some of the basic features and relationships in the tournaments studied here, Section I.2 of the present paper reproduces a subset of the results in Gross (2015b). In particular, Table I.4 uses the companion paper's sample to estimate the relationship between contest characteristics such as the prize, difficulty, and frequency of feedback and key outcomes, and Table I.5 uses it to estimate the relationship between a design's rating and its probability of being selected as the winner. I invoke the Gross (2015b) sample in these cases because they require a large sample of contests to obtain precise, consistent estimates, and because they are broadly similar across a large set of observable characteristics, as shown below.

The dataset in Gross (2015b) 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 growth had begun to stabilize.

Table H.1 compares characteristics of contests in the two samples. The contests in the Gross (2015b) 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 (2015b) data. Interestingly, although the total number of designs is on average higher in the Gross (2015b) 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 (2015b) sample period. Interestingly, the fraction of contests awarded is nevertheless nearly the same in the two samples.

Tables H.2 and H.3 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 the previous table.

Table H.1: Comparing Samples: Contest characteristics

| | **Gross (2015b)** | **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 H.2: Comparing Samples: Distribution of ratings

| | **Gross (2015b)** | **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 H.3: Comparing Samples: Design batches by size of batch

| | **Gross (2015b)** | **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 |

# Appendix for Chapter II

## II.A Proofs of Propositions

**P**roposition 1: The returns to a player's second design decline as the quality of her first design and the cumulative competition grow distant, approaching zero at the limit. Feedback that reveals these asymmetries will therefore discourage participation, relative to a state of ignorance.

**P**roof:

The discouraging effects of asymmetries are a standard result in the tournament literature. Formally:

**C**ase 1: Player with a low first draw ($\beta^L$)

The expected increase in the player's probability of winning from a second draw is:

$$q\left(\frac{\beta^L+\beta^H}{\beta^L+\beta^H+\mu}\right)+(1-q)\left(\frac{2\beta^L}{2\beta^L+\mu}\right)-\left(\frac{\beta^L}{\beta^L+\mu}\right)$$

As $\mu \longrightarrow 0$, this quantity approaches $q\left(\frac{\beta^L+\beta^H}{\beta^L+\beta^H}\right)+(1-q)\left(\frac{2\beta^L}{2\beta^L}\right)-\left(\frac{\beta^L}{\beta^L}\right)=0$

As $\mu \longrightarrow \infty$, this quantity approaches $\frac{1}{\mu}\left(q\left(\beta^L+\beta^H\right)+(1-q)\left(2\beta^L\right)-\beta^L\right) \longrightarrow 0$

**C**ase 2: Player with a high first draw ($\beta^H$)

The expected increase in the player's probability of winning from a second draw is:

$$\left(\frac{2\beta^H}{2\beta^H+\mu}\right)-\left(\frac{\beta^H}{\beta^H+\mu}\right)$$

As $\mu \longrightarrow 0$, this quantity approaches $\left(\frac{2\beta^H}{2\beta^H}\right)-\left(\frac{\beta^H}{\beta^H}\right)=1-1=0$

As $\mu \longrightarrow \infty$, this quantity approaches $\frac{1}{\mu}\left(2\beta^H-\beta^H\right)={}^{\beta^H}/_{\mu} \longrightarrow 0$

**P**roposition 2: Upon provision of feedback, provided competition is sufficiently high, players with better feedback have higher incentives to participate than those with lower feedback.

**P**roof:

The proposition asserts that provided $\mu$ is sufficiently high, players with positive feedback benefit more from their second design than players with negative feedback.

In mathematical notation, the claim is that there exists a $\mu^*$ such that when $\mu > \mu^*$,

$$\underbrace{\left(\frac{2\beta^H}{2\beta^H+\mu}\right)-\left(\frac{\beta^H}{\beta^H+\mu}\right)}_{\text{Benefit to effort}|\text{pos. feedback}} > \underbrace{q\left(\frac{\beta^L+\beta^H}{\beta^L+\beta^H+\mu}\right)+(1-q)\left(\frac{2\beta^L}{2\beta^L+\mu}\right)-\left(\frac{\beta^L}{\beta^L+\mu}\right)}_{\text{Benefit to effort}|\text{neg. feedback}}$$

or in other words, that there exists a $\mu^*$ such that for all $\mu > \mu^*$,

$$\left[\left(\frac{2\beta^H}{2\beta^H+\mu}\right)-q\left(\frac{\beta^L+\beta^H}{\beta^L+\beta^H+\mu}\right)-(1-q)\left(\frac{2\beta^L}{2\beta^L+\mu}\right)\right]-\left[\left(\frac{\beta^H}{\beta^H+\mu}\right)-\left(\frac{\beta^L}{\beta^L+\mu}\right)\right]>0$$

To support this claim, I derive the shape of the expression above and show that it is always positive beyond some fixed, implicitly-defined level of competition. First, note that as $\mu \longrightarrow 0$ or $\mu \longrightarrow \infty$, the expression goes to 0, by the same arguments as in the proof to Proposition 1.

The derivative of the expression is

$$\begin{aligned}&\frac{\partial}{\partial\mu}\left[\left(\frac{2\beta^H}{2\beta^H+\mu}\right)-q\left(\frac{\beta^L+\beta^H}{\beta^L+\beta^H+\mu}\right)-(1-q)\left(\frac{2\beta^L}{2\beta^L+\mu}\right)\right]-\left[\left(\frac{\beta^H}{\beta^H+\mu}\right)-\left(\frac{\beta^L}{\beta^L+\mu}\right)\right]\\&=\left[\left(\frac{-2\beta^H}{(2\beta^H+\mu)^2}\right)-q\left(\frac{-\beta^L-\beta^H}{(\beta^L+\beta^H+\mu)^2}\right)-(1-q)\left(\frac{-2\beta^L}{(2\beta^L+\mu)^2}\right)\right]-\left[\left(\frac{-\beta^H}{(\beta^H+\mu)^2}\right)-\left(\frac{-\beta^L}{(\beta^L+\mu)^2}\right)\right]\end{aligned}$$

As $\mu \longrightarrow 0$, the derivative goes to

$$\begin{aligned}&\left[q\left(\frac{1}{\beta^L+\beta^H}\right)+(1-q)\left(\frac{1}{2\beta^L}\right)-\left(\frac{1}{2\beta^H}\right)\right]-\left[\left(\frac{1}{\beta^L}\right)-\left(\frac{1}{\beta^H}\right)\right]\\&=\left[q\left(\frac{1}{\beta^L+\beta^H}\right)+(1-q)\left(\frac{1}{2\beta^L}\right)-\frac{1}{2\beta^H}\right]-\left[\frac{1}{\beta^L}-\frac{1}{\beta^H}\right]\\&<\left[q\left(\frac{1}{\beta^L+\beta^L}\right)+(1-q)\left(\frac{1}{2\beta^L}\right)-\frac{1}{2\beta^H}\right]-\left[\frac{1}{\beta^L}-\frac{1}{\beta^H}\right]\\&=\frac{1}{2}\left[\frac{1}{\beta^L}-\frac{1}{\beta^H}\right]-\left[\frac{1}{\beta^L}-\frac{1}{\beta^H}\right]=-\frac{1}{2}\left[\frac{1}{\beta^L}-\frac{1}{\beta^H}\right]<0\text{ , since }\beta^H>\beta^L\end{aligned}$$

As $\mu \longrightarrow \infty$, the derivative goes to

$$\begin{aligned}&\frac{1}{\mu^2}\left[q\left(\beta^L+\beta^H\right)+(1-q)\left(2\beta^L\right)-2\beta^H+\beta^H-\beta^L\right]\\&=\frac{1}{\mu^2}\left[q\left(\beta^L+\beta^H\right)+(1-q)\left(2\beta^L\right)-\beta^H-\beta^L\right]\\&=\frac{1}{\mu^2}\left[q\beta^H-q\beta^L-\left(\beta^H-\beta^L\right)\right]\\&=\frac{1}{\mu^2}\left[(q-1)\left(\beta^H-\beta^L\right)\right]\longrightarrow 0^-\end{aligned}$$

Additionally, the expression can be shown to have at most three roots, and only one positive, real root. To demonstrate this, set the expression equal to zero:

$$\left[\left(\frac{2\beta^H}{2\beta^H+\mu}\right) - q\left(\frac{\beta^L+\beta^H}{\beta^L+\beta^H+\mu}\right) - (1-q)\left(\frac{2\beta^L}{2\beta^L+\mu}\right)\right] - \left[\left(\frac{\beta^H}{\beta^H+\mu}\right) - \left(\frac{\beta^L}{\beta^L+\mu}\right)\right] = 0$$

To simplify the notation, redefine $H = \beta^H$ and $L = \beta^L$. Additionally, let

$$\begin{aligned}
A &= 2H + \mu \\
B &= H + L + \mu \\
C &= 2L + \mu \\
D &= H + \mu \\
E &= L + \mu
\end{aligned}$$

We can then rewrite the equation as the following:

$$[2H \cdot BCDE] - [q(H+L) \cdot ACDE] - [(1-q)(2L) \cdot ABDE] - [H \cdot ABCE] + [L \cdot ABCD] = 0$$

Rearranging more terms, we get:

$$[H \cdot BCE\,(2D - A)] - [L \cdot ABD\,(2E - C)] + [q \cdot ADE\,((2B - C)\,L - (C)\,H)] = 0$$

Observe that $2D - A = 2E - C = \mu$, and that $2B - C = A$. Then simplifying further:

$$\mu\,[H \cdot BCE] - \mu\,[L \cdot ABD] + qADE\,(AL - CH) = 0$$

Now observe that $AL - CH = \mu(L - H)$. We continue simplifying:

$$\begin{aligned}
\mu\,[H \cdot BCE] - \mu\,[L \cdot ABD] + \mu\,[qADE\,(L - H)] &= 0 \\
[H \cdot BCE] - [L \cdot ABD] + [qADE\,(L - H)] &= 0 \\
[H \cdot BCE] - [L \cdot ABD] - [qADE\,(H - L)] &= 0
\end{aligned}$$

which is now cubic in $\mu$ (reduced by algebra from what was ostensibly quartic).

Additionally, it can be shown that

$$BCE = \mu^3 + (H + 4L)\,\mu^2 + \left(3HL + 5L^2\right) + \left(2HL^2 + 2L^3\right)$$

and by symmetry,

$$ABD = \mu^3 + (4H + L)\,\mu^2 + \left(3HL + 5H^2\right) + \left(2H^2L + 2H^3\right)$$

such that

$$\begin{aligned}
H \cdot BCE &= H\mu^3 + \left(H^2 + 4HL\right)\mu^2 + \left(3H^2L + 5HL^2\right) + \left(2H^2L^2 + 2HL^3\right) \\
L \cdot ABD &= L\mu^3 + \left(4HL + L^2\right)\mu^2 + \left(3HL^2 + 5H^2L\right) + \left(2H^2L^2 + 2H^3L\right)
\end{aligned}$$

and the difference between them being

$$\begin{aligned}
H \cdot BCE - L \cdot ABD &= (H-L)\,\mu^3 + \left(H^2 - L^2\right)\mu^2 + HL\,(3(H-L) + 5(L-H))\,\mu + 2HL\left(L^2 - H^2\right) \\
&= (H-L)\,\mu^3 + \left(H^2 - L^2\right)\mu^2 + HL\,(3(H-L) - 5(H-L))\,\mu + 2HL\left(L^2 - H^2\right) \\
&= (H-L)\left[\mu^3 + (H+L)\,\mu^2 - 2HL\mu - 2HL\,(H+L)\right]
\end{aligned}$$

Now note that

$$ADE = \mu^3 + (3H+L)\,\mu^2 + \left(3HL + 2H^2\right)\mu + 2H^2L$$

Returning to above, we can then write

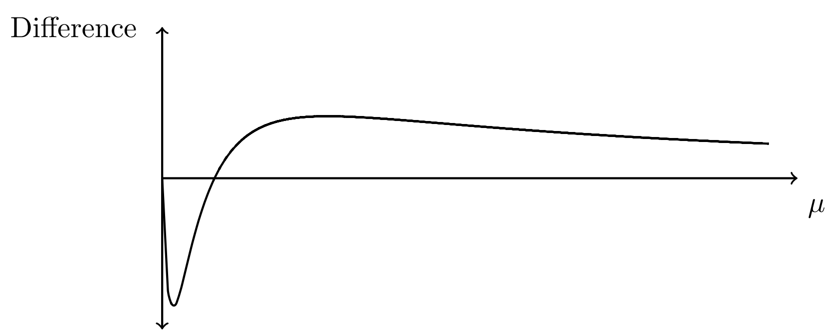
$$\begin{aligned}
[H \cdot BCE] - [L \cdot ABD] - [qADE\,(H-L)] = &\left[\mu^3 + (H+L)\,\mu^2 - 2HL\mu - 2HL\,(H+L)\right] \\
&- q\left[\mu^3 + (3H+L)\,\mu^2 + \left(3HL + 2H^2\right)\mu + 2H^2L\right] = 0
\end{aligned}$$

and in the final simplification,

$$\begin{aligned}
&(1-q)\,\mu^3 \\
&+ ((1-3q)\,H + (1-q)\,L)\,\mu^2 \\
&-H\,(2qH + 3qL + 2L)\,\mu \\
&-2HL\,((1+q)\,H + L) \quad = 0
\end{aligned}$$

By the rule of signs, the polynomial has exactly one positive, real root.

The difference between the benefits to a second design after positive feedback relative to the benefits after negative feedback is thus (i) zero with no competition, (ii) decreasing and negative as competition initially grows, (iii) later increasing and turning positive, and (iv) eventually decreasing and asymptoting towards zero as competition grows infinitely large, as in the figure below. Beyond a fixed, intermediate (and typically relatively low) $\mu$, this difference will therefore always be positive.



## II.B Dataset Construction

The sample for this paper consists of all publicly viewable logo design contests on the platform from July 2010 through June 2012 with a single, winner-take-all prize and no mid-contest rule changes. Although the complete dataset includes contests in other design categories dating back to the platform's inception, logo design is the modal contest category and is thus a natural choice for analysis. The focal sample begins over two years after the platform was created, by which point it was a well-known, established player in the graphic design industry, most of its features were set, and its growth had begun to stabilize.

### 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: his/her experience in previous contests on the platform (number of contests and designs entered, contests won, prize winnings, recent activity), average ratings from previous contests, 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.[3] In contrast to Gross (2014), I do not have the designs themselves for this sample.

The full dataset – most of which is not used in this paper – consists of nearly all contests with public bidding completed since the birth of the platform in 2008, or about 80 percent of all contests on the platform through August 1, 2012. I use these contests to re-construct players' history on the platform up to each contest that they enter in my primary sample and over the entire four-year period.

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.

[3]Note that the "withdrawn" indicator is unreliable, as all of a user's designs are flagged as withdrawn whenever the user deletes her account from the website – including designs in completed contests. In the analysis, I assume withdrawn entries remain in contention, as sponsors can ask for any withdrawn design to be reinstated.

## II.C Additional Tables

### Abandonment after Second Rating

This appendix provides robustness checks on the results of Section II.3, which evaluates the effects of a player's first rating on her continued participation, by presenting analogous tables for second ratings.

Table II.C.1 is the counterpart to Table II.6 from the paper, showing the distribution of the number of designs a player enters after her second rating, conditional on that rating being one star or four to five stars. While the majority (63.6 percent) of players receiving a 1-star rating will drop out, the majority (60.0 percent) of those receiving a 4- or 5-star rating will enter more designs. For comparison, recall that the analogous frequencies for first ratings were 69.5 percent and 61.2 percent, respectively.

Table II.C.1: Designs entered after a player's second rating

| **When second rating is 1 (out of 5)** | | | **When second rating is 4 or 5 (out of 5)** | | |
|---|---|---|---|---|---|
| **Designs after 2nd rating** | **Freq.** | **Pct.** | **Designs after 2nd rating** | **Freq.** | **Pct.** |
| 0 designs | 4,891 | 63.60 | 0 designs | 5,329 | 39.99 |
| 1 design | 764 | 9.93 | 1 design | 1,803 | 13.53 |
| 2 designs | 584 | 7.59 | 2 designs | 1,574 | 11.81 |
| 3 designs | 334 | 4.34 | 3 designs | 1,113 | 8.35 |
| 4 designs | 257 | 3.34 | 4 designs | 899 | 6.75 |
| 5 designs | 213 | 2.77 | 5 designs | 605 | 4.54 |
| 6 designs | 133 | 1.73 | 6 designs | 441 | 3.31 |
| 7 designs | 93 | 1.21 | 7 designs | 316 | 2.37 |
| 8 designs | 69 | 0.90 | 8 designs | 239 | 1.79 |
| 9 designs | 73 | 0.95 | 9 designs | 221 | 1.66 |
| 10+ designs | 279 | 3.63 | 10+ designs | 785 | 5.89 |
| **Total** | **7,690** | **100** | **Total** | **13,325** | **100** |

Notes: Table reports the activity of players after receiving their second rating in a contest, by the value of that second rating. A total of 50,229 contest-players received second ratings. Of these: 7,690 were rated 1 star (15.3 percent); 12,182 were rated 2 stars (24.3 percent); 17,032 were rated 3 stars (33.9 percent); 11,043 were rated 4-star (22.0 percent); and 2,282 were rated 5-star (4.5 percent). The table illustrates that players are much more likely to continue participating in a contest after positive feedback, similar to the results for first ratings.

Table II.C.2 formalizes these results, as in Table II.7 for the first rating. Recall the specification:

$$\begin{aligned} Abandon_{jk} &= \beta_0 + \sum_{r=1}^{5} \beta_r \cdot \mathbb{1}(R_{jk} = r) + \sum_{r=1}^{5} \gamma_r \cdot \mathbb{1}(\bar{R}_{-jk} = r) \\ &\quad + \delta \cdot Timing_{jk} + X_{jk}\theta + \zeta_k + \varphi_j + \varepsilon_{jk} \end{aligned}$$

where $Abandon_{jk}$ now indicates that player $j$ entered no designs in contest $k$ after her $s$econd rating; $R_{jk}$ is the player's second rating; $\bar{R}_{ijk}$ is the highest rating on any competing designs at that time; $Timing_{jk}$ is the fraction of the contest elapsed at the time of that rating; $X_{jk}$ is a vector of controls; and $\zeta_k$ and $\varphi_j$ are contest and player fixed effects, respectively. The table provides the same sequence of specifications presented the body of the paper: linear models with contest, player, and contest and player fixed effects in Columns (1) to (3), and a logit model with contest fixed effects in Column (4).

Table II.C.2: Tendency to abandon after second rating, as function of rating

| | Dependent variable: Abandon after second rating | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linear | Linear | Linear | Logit |
| Player's second rating==5 | -0.266*** | -0.265*** | -0.316*** | -1.408*** |
| | (0.024) | (0.029) | (0.031) | (0.119) |
| Player's second rating==4 | -0.266*** | -0.255*** | -0.309*** | -1.410*** |
| | (0.016) | (0.018) | (0.020) | (0.078) |
| Player's second rating==3 | -0.191*** | -0.173*** | -0.211*** | -1.019*** |
| | (0.014) | (0.016) | (0.017) | (0.069) |
| Player's second rating==2 | -0.096*** | -0.083*** | -0.112*** | -0.516*** |
| | (0.013) | (0.015) | (0.016) | (0.064) |
| Competitors' prior best==5 | 0.039 | 0.069*** | 0.054* | 0.323** |
| | (0.024) | (0.021) | (0.029) | (0.127) |
| Constant | 0.420*** | 0.365*** | -0.383* | -1.053 |
| | (0.039) | (0.034) | (0.203) | (0.735) |
| N | 25996 | 25996 | 25996 | 24139 |
| $R^2$ | 0.27 | 0.38 | 0.51 | |
| Contest FEs | Yes | No | Yes | Yes |
| Player FEs | No | Yes | Yes | No |

Notes: Table shows the effect of a player's second rating in a contest and the competition at that time on the probability that the player subsequently enters no more designs. Observations are contest-players. The dependent variable in all columns is an indicator for whether the player abandons after her second rating. Columns (1) to (3) estimate linear models with fixed effects; Column (4) estimates a logit model without player fixed effects, which may render the estimates inconsistent. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. All specifications control for the player's first rating and the time remaining, both in levels and as a percent of the contest duration. Standard errors clustered by player in parentheses.

The effect of second ratings on the probability that a player enters more designs is similar to those of first ratings, albeit a bit attenuated at the top. However, players with more positive feedback are again significantly more likely to remain active than those with poor initial feedback, and high-rated competition continues to make it more likely that a player abandons.

Table II.C.3 reestimates the model above as a quadratic in the probability of winning after the second rating, as in Table II.8 for first ratings. The tendency to abandon remains significantly convex in a player's probability of winning, and is still minimized near a win probability of 50 percent.

Table II.C.3: Tendency to abandon after second rating, as function of Pr(Win)

| | Dependent variable: Abandon after 2nd rating | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| | Linear | Linear | Linear | Logit |
| Pr(Win) | -1.208*** | -1.195*** | -1.311*** | -6.698*** |
| | (0.058) | (0.065) | (0.074) | (0.332) |
| Pr(Win)$^2$ | 1.109*** | 1.068*** | 1.199*** | 5.955*** |
| | (0.069) | (0.076) | (0.086) | (0.396) |
| Constant | 0.345*** | 0.298*** | -0.554*** | -1.118 |
| | (0.036) | (0.028) | (0.207) | (0.763) |
| N | 25996 | 25996 | 25996 | 24139 |
| $R^2$ | 0.26 | 0.37 | 0.50 | |
| Contest FEs | Yes | No | Yes | Yes |
| Player FEs | No | Yes | Yes | No |
| Minimizer | 0.54 | 0.56 | 0.55 | 0.56 |

Notes: Table shows the effect of a player's win probability after receiving her second rating on the probability that she subsequently enters no more designs. Observations are contest-players. The dependent variable in all columns is an indicator for whether the player abandons after her second rating. Columns (1) to (3) estimate linear models with fixed effects; Column (4) estimates a logit model without player fixed effects, which may render the estimates inconsistent. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. All specifications control for the time remaining, both in levels and as a percent of the contest duration. Standard errors clustered by player in parentheses.

### Conditional Logit Estimates

Table II.C.4 provides the conditional logit estimates for the success function in Section II.4. Recall that the latent value of a design is modeled as linear in fixed effects for each rating and an i.i.d. Type-I E.V. error:

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

The table provides estimates for the $\gamma$s and is discussed in greater detail in the text. The content of the discussion is copied below for reference:

> Several patterns emerge [from Table II.C.4]. First, the value of a design is monotonically increasing in its rating, with only a 5-star rating being on average preferred to the outside option, and the fixed effects are precisely estimated. To produce the same change in the success function generated by a 5-star design, a player would need 12 4-star designs, 137 3-star designs, or nearly 2,000 1-star designs – so competition effectively comes from the top. As a measure of fit, the predicted odds-on favorite wins almost half of all contests in the sample. These results demonstrate that this simple model fits the data quite well and in an intuitive way, suggesting that ratings provide considerable information about a player's probability of winning.

Table II.C.4: Conditional logit of win-lose outcomes on ratings

| **Fixed effect** | **Est.** | **S.E.** | **t-stat** |
|---|---|---|---|
| Rating==5 | 1.53 | 0.07 | 22.17 |
| Rating==4 | -0.96 | 0.06 | -15.35 |
| Rating==3 | -3.39 | 0.08 | -40.01 |
| Rating==2 | -5.20 | 0.17 | -30.16 |
| Rating==1 | -6.02 | 0.28 | -21.82 |
| No rating | -3.43 | 0.06 | -55.35 |

Notes: Table provides estimates from conditional logit estimation of the win-lose outcome of each design as a function of its rating. Outside option is not awarding the prize, with utility normalized to zero. The design predicted by the model as the odds-on favorite wins roughly 50 percent of contests.

The results of this exercise make it possible to compute predicted probabilities of winning for any player at any ratings history, and are used towards this end in several parts of the paper, including regressions, estimation of effort costs, and simulations.

### Evidence that costs are not mechanical

Table II.C.5 regresses the log probability gains achieved by a player's final design or an extra design on the log prize. The estimates reveal that large prizes are competed away: when the prize increases by one percent, the probability gains of players' final submissions declines by 0.75 percent, with the remaining 0.25 percent being reflected in a higher cost. This evidence supports the argument that the estimated costs are meaningful rather than mechanical. See Section II.5 for further discussion.

Table II.C.5: Evidence that players compete away prize increases

| | Log of win probability gains acheived by: | |
|---|---|---|
| | Final design | Extra design |
| Log prize | -0.752*** | -0.734*** |
| | (0.006) | (0.006) |
| Constant | -0.047 | -0.180*** |
| | (0.034) | (0.034) |
| N | 160059 | 160059 |
| $R^2$ | 0.10 | 0.10 |

Notes: Table shows the correlation between the prize and the estimated probability gains (1) achieved by players' final designs, and (2) available from players' next, unentered designs. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

## II.D Evidence of Foresight

The structural model in Section II.4 requires an assumption that at the time of their final submission, players can predict the level of competition they will face at the end of a contest. This assumption is necessary for players to be able to compute the benefits to continued participation at the time of exit, which I use to bound their cost. This appendix provides empirical support for the assumption.

Two pieces of information are required to support this claim: the distribution of exit times and the quality of predictions made at those times. Figure II.D.1 provides the cumulative distribution of a player's first and last designs in a contest, calculated over the fraction of the contest elapsed. The figure shows that players tend to exit in the later stages of a contest: roughly half of all contest-players exit a given contest after 80 percent of the contest has transpired, and nearly two-thirds exit after 50 percent has transpired.

Given that players tend to exit late, the question remains as to how well they can forecast the terminal state of competition at that time. Table II.D.1 shows the goodness-of-fit ($R^2$) of a regression of the terminal number of competing designs of each rating on the number observed after a given fraction of the contest has elapsed. This method can predict the total number of competing designs with a high degree of accuracy ($R^2 = 0.88$) when only half of the contest has elapsed, and even better ($R^2 = 0.97$) when 80 percent of the contest has elapsed. Given that competition tends to come from the top, we may be more interested in the quality of forecasts over top-rated competitors: predictions of the terminal number of 5-star designs at the 50 percent mark and 80 percent mark have an $R^2 = 0.67$ and $R^2 = 0.90$. Figures II.D.2 to II.D.8 provide scatterplots of the regressions in Table II.D.1, so that the goodness-of-fit can be visualized.

The combined evidence that the majority of players exit in the latter half of a contest and that terminal levels of competition can be forecast from the levels observed at that time provide support to the assumption of foresight in the empirical model.

Table II.D.1: Predictability of final number of competing designs of each rating

| **Percent of contest elapsed** | **All** | **Unrated** | **1-star** | **2-star** | **3-star** | **4-star** | **5-star** |
|---|---|---|---|---|---|---|---|
| 50 | 0.88 | 0.66 | 0.37 | 0.68 | 0.69 | 0.55 | 0.67 |
| 51 | 0.86 | 0.74 | 0.37 | 0.65 | 0.75 | 0.62 | 0.67 |
| 52 | 0.87 | 0.78 | 0.48 | 0.71 | 0.78 | 0.65 | 0.71 |
| 53 | 0.89 | 0.74 | 0.37 | 0.68 | 0.80 | 0.63 | 0.75 |
| 54 | 0.90 | 0.75 | 0.39 | 0.75 | 0.78 | 0.54 | 0.68 |
| 55 | 0.90 | 0.80 | 0.51 | 0.73 | 0.78 | 0.62 | 0.78 |
| 56 | 0.91 | 0.75 | 0.54 | 0.74 | 0.80 | 0.58 | 0.74 |
| 57 | 0.91 | 0.72 | 0.43 | 0.71 | 0.78 | 0.60 | 0.75 |
| 58 | 0.92 | 0.78 | 0.56 | 0.77 | 0.87 | 0.83 | 0.77 |
| 59 | 0.90 | 0.81 | 0.62 | 0.76 | 0.84 | 0.83 | 0.76 |
| 60 | 0.91 | 0.77 | 0.50 | 0.75 | 0.83 | 0.84 | 0.82 |
| 61 | 0.93 | 0.85 | 0.63 | 0.79 | 0.86 | 0.86 | 0.77 |
| 62 | 0.92 | 0.81 | 0.52 | 0.75 | 0.79 | 0.78 | 0.74 |
| 63 | 0.94 | 0.76 | 0.48 | 0.81 | 0.88 | 0.86 | 0.78 |
| 64 | 0.93 | 0.73 | 0.48 | 0.78 | 0.86 | 0.85 | 0.76 |
| 65 | 0.94 | 0.82 | 0.54 | 0.79 | 0.86 | 0.87 | 0.78 |
| 66 | 0.94 | 0.84 | 0.52 | 0.84 | 0.88 | 0.88 | 0.79 |
| 67 | 0.93 | 0.76 | 0.44 | 0.79 | 0.85 | 0.83 | 0.77 |
| 68 | 0.94 | 0.80 | 0.48 | 0.78 | 0.87 | 0.86 | 0.77 |
| 69 | 0.94 | 0.69 | 0.40 | 0.81 | 0.87 | 0.85 | 0.80 |

Table II.D.1: Predictability of final number of competing designs of each rating (cont'd)

| **Percent of contest elapsed** | **All** | **Unrated** | **1-star** | **2-star** | **3-star** | **4-star** | **5-star** |
|---|---|---|---|---|---|---|---|
| 70 | 0.95 | 0.76 | 0.42 | 0.82 | 0.87 | 0.85 | 0.73 |
| 71 | 0.94 | 0.76 | 0.44 | 0.78 | 0.81 | 0.82 | 0.77 |
| 72 | 0.94 | 0.79 | 0.51 | 0.83 | 0.86 | 0.84 | 0.83 |
| 73 | 0.95 | 0.82 | 0.44 | 0.83 | 0.87 | 0.85 | 0.81 |
| 74 | 0.95 | 0.82 | 0.46 | 0.80 | 0.88 | 0.88 | 0.82 |
| 75 | 0.95 | 0.76 | 0.52 | 0.78 | 0.88 | 0.87 | 0.79 |
| 76 | 0.96 | 0.70 | 0.45 | 0.80 | 0.87 | 0.86 | 0.77 |
| 77 | 0.96 | 0.76 | 0.39 | 0.85 | 0.86 | 0.87 | 0.80 |
| 78 | 0.97 | 0.86 | 0.48 | 0.84 | 0.88 | 0.86 | 0.82 |
| 79 | 0.97 | 0.83 | 0.58 | 0.90 | 0.91 | 0.90 | 0.86 |
| 80 | 0.97 | 0.83 | 0.85 | 0.88 | 0.89 | 0.90 | 0.90 |
| 81 | 0.97 | 0.88 | 0.92 | 0.85 | 0.88 | 0.89 | 0.89 |
| 82 | 0.97 | 0.91 | 0.95 | 0.87 | 0.88 | 0.87 | 0.86 |
| 83 | 0.97 | 0.89 | 0.92 | 0.88 | 0.90 | 0.87 | 0.88 |
| 84 | 0.97 | 0.89 | 0.90 | 0.85 | 0.90 | 0.88 | 0.88 |
| 85 | 0.97 | 0.92 | 0.93 | 0.83 | 0.87 | 0.87 | 0.90 |
| 86 | 0.98 | 0.91 | 0.84 | 0.82 | 0.88 | 0.87 | 0.87 |
| 87 | 0.98 | 0.93 | 0.91 | 0.85 | 0.92 | 0.89 | 0.92 |
| 88 | 0.98 | 0.93 | 0.95 | 0.87 | 0.93 | 0.90 | 0.90 |
| 89 | 0.98 | 0.95 | 0.94 | 0.90 | 0.93 | 0.89 | 0.92 |
| 90 | 0.98 | 0.94 | 0.97 | 0.88 | 0.92 | 0.88 | 0.90 |
| 91 | 0.99 | 0.93 | 0.95 | 0.87 | 0.92 | 0.88 | 0.90 |
| 92 | 0.99 | 0.92 | 0.96 | 0.90 | 0.92 | 0.89 | 0.89 |
| 93 | 0.99 | 0.92 | 0.97 | 0.90 | 0.92 | 0.88 | 0.91 |
| 94 | 0.99 | 0.92 | 0.97 | 0.92 | 0.92 | 0.89 | 0.93 |
| 95 | 0.99 | 0.95 | 0.99 | 0.92 | 0.94 | 0.91 | 0.95 |
| 96 | 0.99 | 0.94 | 0.99 | 0.91 | 0.94 | 0.89 | 0.94 |
| 97 | 0.99 | 0.95 | 0.99 | 0.93 | 0.94 | 0.89 | 0.94 |
| 98 | 1.00 | 0.95 | 0.98 | 0.94 | 0.94 | 0.91 | 0.95 |
| 99 | 1.00 | 0.96 | 0.99 | 0.96 | 0.96 | 0.89 | 0.95 |

Notes: Table provides $R^2$ from regressions of the final number of competing designs of each rating on the number observed after a given fraction of the contest has elapsed. Observations are individual submissions; for each submission I record the number of competing designs at that time and seek to project the state of competition when the contest concludes. The high fit suggests that future levels of competition (especially top-rated competition) can be reasonably well forecast in the latter half of any contest, when the majority of players stop competing (Figure II.D.1), supporting the modeling assumption that players have foresight.

Figure II.D.1: Empirical CDFs of player entry and exit

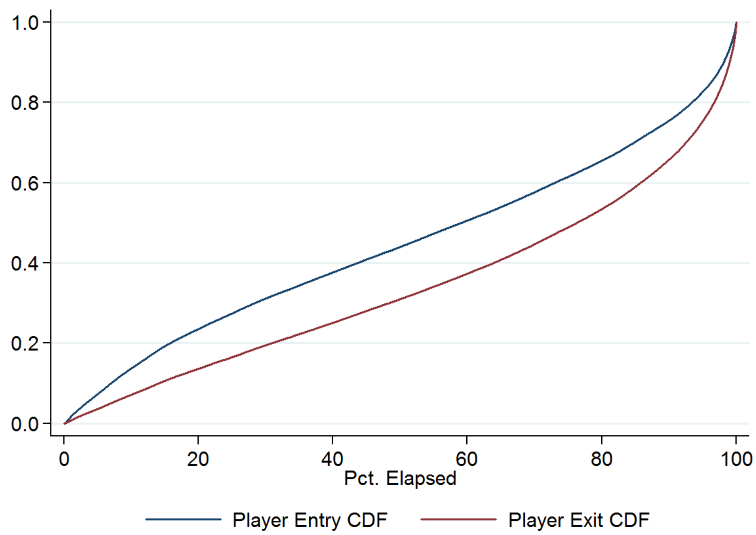



Figure II.D.2: Predictability of final number of competing designs

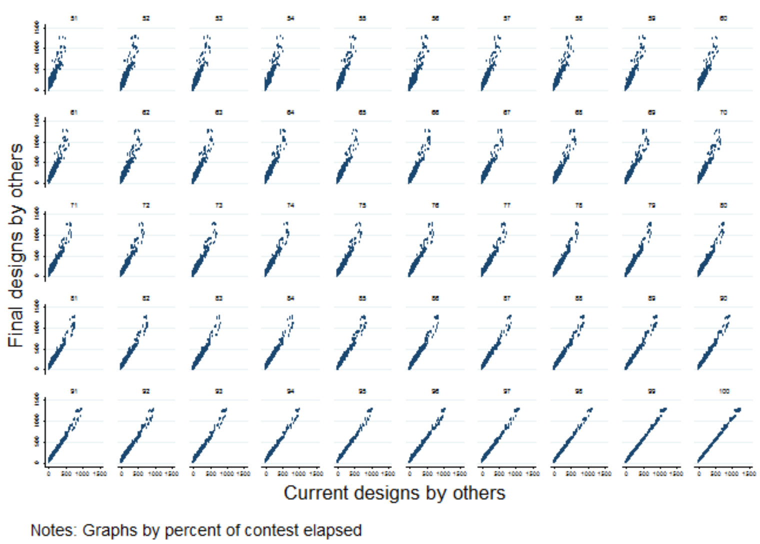
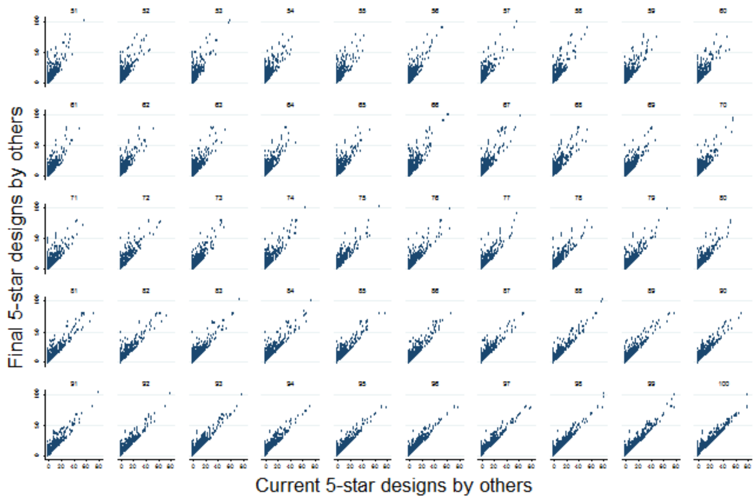
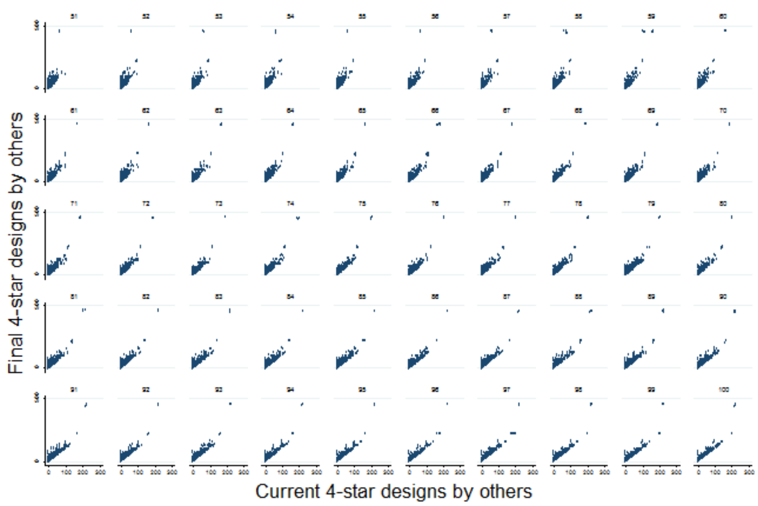


Figure II.D.3: Predictability of final number of competing 5-star designs




Notes: Graphs by percent of contest elapsed

Figure II.D.4: Predictability of final number of competing 4-star designs




Notes: Graphs by percent of contest elapsed

Figure II.D.5: Predictability of final number of competing 3-star designs

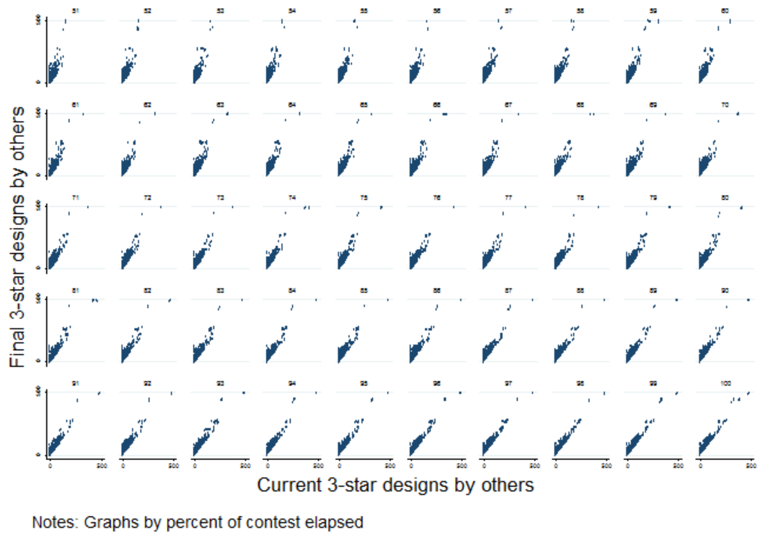



Figure II.D.6: Predictability of final number of competing 2-star designs

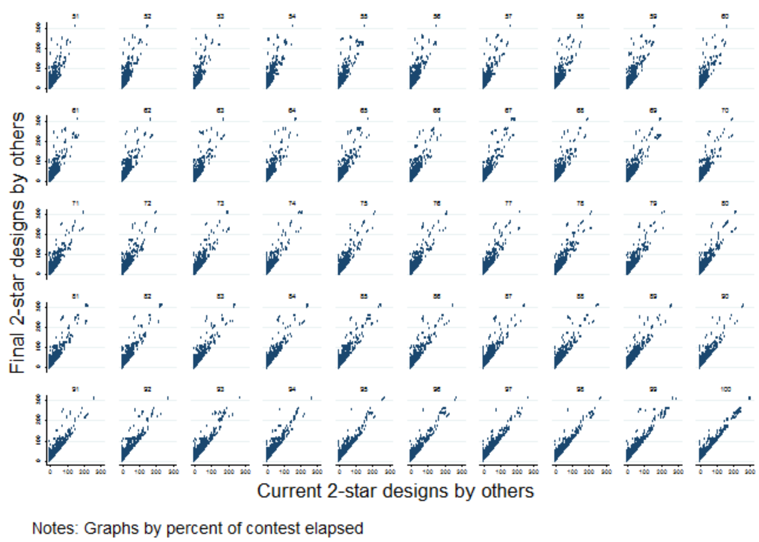


Figure II.D.7: Predictability of final number of competing 1-star designs

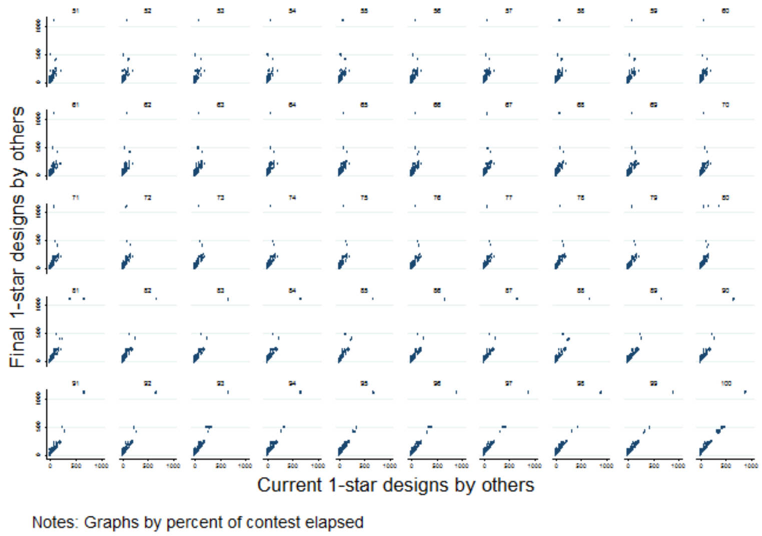
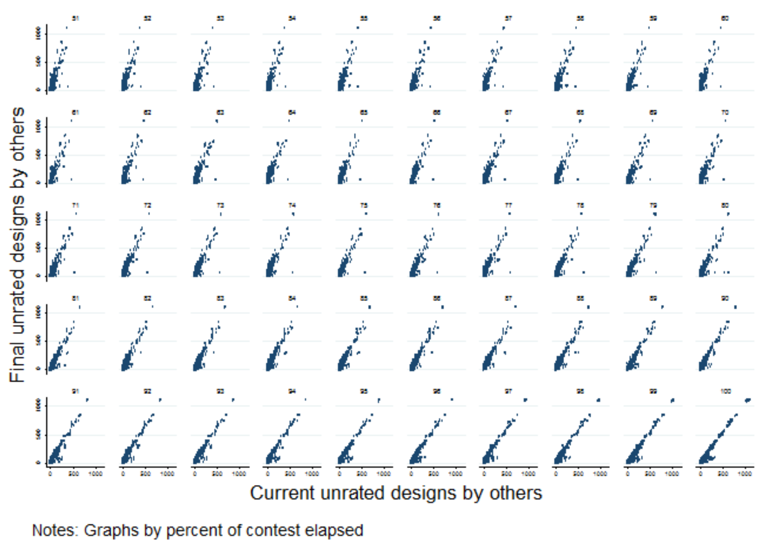



Figure II.D.8: Predictability of final number of competing unrated designs



# Appendix for Chapter III

## III.A Data Appendix

The majority of the data used in the paper come from the U.S. Census of Agriculture for years 1910, 1920, 1925, 1930, and 1940. When possible, data were acquired from NHGIS; remaining variables were encoded from PDF files obtained from the Census website.[4] Stock variables (e.g., farms, farmland, number and value of farm machinery and draft animals, etc.) are reported for the Census year; flows (inputs, outputs) are always from the preceding year. Where corn acreage is separately reported for corn harvested for grain, cut for silage, cut for fodder, and hogged or grazed off (1925 and later), I use the acreage of corn harvested for grain, which is typically around 90 percent of total corn acreage and the subset most relevant to mechanization. Certain crops are not reported for certain states in certain years (barley and rye in Missouri, rye in Kansas – both in 1930) due to omission from the state-specific questionnaire, which likely resulted from low acreage; production of these crops in the affected counties is coded as zero. Occasionally, a page went missing in the Census documents; in these cases, the affected observations were coded as missing. Scans of the 1954 Census of Agriculture (obtained from the USDA National Agricultural Statistics Service; see link in footnote) provided data on tractor and implement ownership in 1950.

U.S. county shapefiles were obtained from NHGIS for each decade from 1910 to 1940. These maps were used to calculate counties' geographic centroids, mean and standard deviation elevation (calculated from USGS National Elevation Dataset rasters), and average annual rainfall (calculated from the PRISM Climate Group 30-year normals). I use county entry and exit into/out of the dataset and movement in geographic centroids to identify counties that formed, merged, split, or dissolved between Census years; any such counties are dropped from the analysis. As the text explains, I also apply Hornbeck's (2010) county border fix algorithm as a robustness check. I calculate distance to the f.o.b. shipping locations of Ford (Detroit) and International Harvester (Chicago) as a proxy for freight costs; comparison with point-to-point freight rates from Hartman (1916) suggests distance is a reasonable proxy, with correlations between route distance and point-to-point rates of $> 0.95$ for routes originating in Detroit or Chicago.

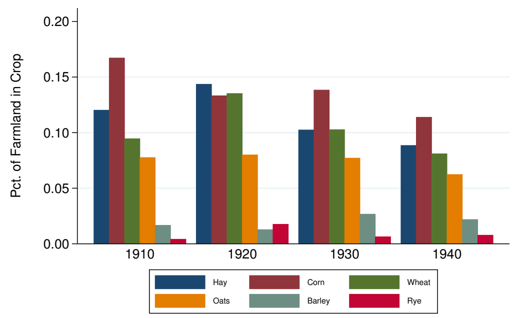
The data used in the New Deal and Dust Bowl robustness checks in Appendix III.F were obtained from Fishback, Kantor, and Wallis (2003) and Hornbeck (2012), respectively. The New Deal robustness checks include the Fishback et al. measures of AAA relief spending and FCA lending by county, normalized by county farm acreage; the Dust Bowl robustness checks uses the Hornbeck measures of low, medium, and high soil erosion. The latter are restricted to Midwest counties for which soil erosion data were available (those in Kansas, Nebraska, North and South Dakota, Iowa, Minnesota).

Hybrid corn diffusion is from the USDA Agricultural Statistics and was provided in digital format by Richard Sutch (Sutch 2011, 2014). Data on water wheels, steam engines, gas engines, and electricity in U.S. manufacturing were obtained from the Atack, Bateman, and Weiss (1999) sample from the manuscript Census of Manufactures for 1850-1870 and the 1910 Census of Manufactures.

[4]Historical Censuses and associated documents are available at http://www.census.gov/prod/www/decennial.html. A complete collection of historical Agricultural Census publications can be found at http://www.agcensus.usda.gov/Publications/Historical_Publications/index.php.

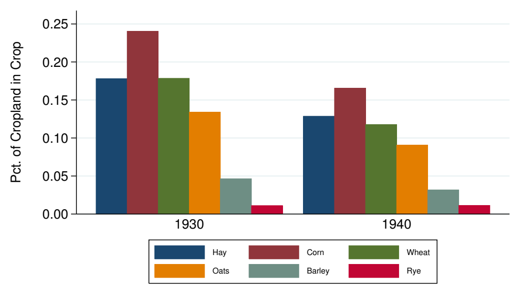
## III.B Additional Descriptive Figures

Figure III.B.1: Crop shares of midwest farmland



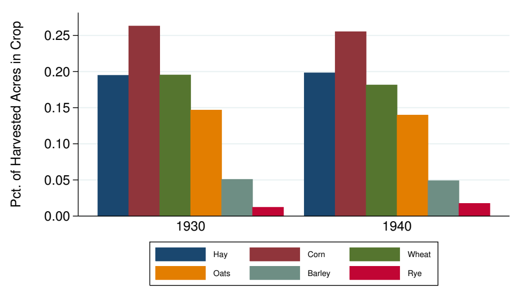

Notes: Percentages are measured as each crop's harvested acreage over total acres of farmland. Corn acreage is limited to corn harvested for grain only (versus silage). The six crops represented here constitute 48% of all acres of farmland in the Midwest in 1930 and 52% in 1940. Data from 1910–1940 Census of Agriculture.

Figure III.B.2: Crop shares of midwest cropland




Notes: Percentages are measured as each crop's harvested acreage over total acres of cropland. Cropland includes harvested crops, crop failure, and land idle or fallow. Corn acreage is limited to corn harvested for grain only (versus fodder or grazed). The six crops represented here constitute 79% of all acres of cropland in the Midwest in 1930 and 55% in 1940. The fraction of cropland in each crop is highly correlated with the fraction of farm–land in each crop, with correlations of 0.94 for corn, 0.97 for wheat, 0.95 for oats, 0.97 for barley, 0.96 for rye, and 0.79 for hay. Data from 1930 and 1940 Census of Agriculture.

Figure III.B.3: Crop shares of midwest harvested acreage




Notes: Percentages are measured as each crop's harvested acreage over total acres of cropland. Corn acreage is limited to corn harvested for grain only (versus fodder or grazed). The six crops represented here constitute 86% of all acres harvested in the Midwest in 1930 and 84% in 1940. The fraction of harvested acreage in each crop is highly correlated with the fraction of farmland in each crop, with correlations of 0.90 for corn, 0.90 for wheat, 0.94 for oats, 0.95 for barley, 0.95 for rye, and 0.73 for hay. Data from 1930 and 1940 Census of Agriculture.

## III.C Regressions for Tractors per Acre

Table III.C.1: Effect of crop mix on tractor diffusion, 1930 and 1940; OLS

| | Diffusion | | Diff. Change | Diffusion | | Diff. Change |
|---|---|---|---|---|---|---|
| | 1930 | 1940 | 1930-1940 | 1930 | 1940 | 1930-1940 |
| Pct. in corn | 0.081** | 0.612*** | 0.302*** | -0.025 | 0.211*** | 0.146*** |
| | (0.040) | (0.061) | (0.028) | (0.049) | (0.054) | (0.030) |
| Pct. in wheat | 0.346*** | 0.497*** | 0.117*** | 0.348*** | 0.330*** | 0.060** |
| | (0.023) | (0.048) | (0.022) | (0.026) | (0.047) | (0.027) |
| Pct. in oats | 0.863*** | 1.152*** | 0.536*** | 0.615*** | 0.931*** | 0.478*** |
| | (0.072) | (0.092) | (0.044) | (0.057) | (0.066) | (0.037) |
| Pct. in barley | 0.018 | 1.075*** | 0.261*** | 0.405*** | 0.841*** | 0.206*** |
| | (0.116) | (0.207) | (0.080) | (0.094) | (0.148) | (0.061) |
| Pct. in rye | -1.257*** | -1.320*** | -0.334*** | -0.734*** | -0.863*** | -0.260*** |
| | (0.230) | (0.293) | (0.099) | (0.185) | (0.231) | (0.093) |
| N | 1034 | 954 | 954 | 1034 | 941 | 941 |
| $R^2$ | 0.50 | 0.63 | 0.65 | 0.75 | 0.87 | 0.82 |
| RMSE | 0.08 | 0.11 | 0.05 | 0.06 | 0.07 | 0.04 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor in 1930, 1940, and from 1930-1940. All specifications regress tractors per 100 farm acres on contemporaneous crop mixes. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.C.2: Effect of crop mix on tractor diffusion, 1930 and 1940; IV

| | Diffusion | | Diff. Change | Diffusion | | Diff. Change |
|---|---|---|---|---|---|---|
| | 1930 | 1940 | 1930-1940 | 1930 | 1940 | 1930-1940 |
| Pct. in corn | 0.232*** | 0.636*** | 0.371*** | 0.008 | 0.310*** | 0.276*** |
| | (0.054) | (0.083) | (0.037) | (0.064) | (0.099) | (0.050) |
| Pct. in wheat | 0.278*** | 0.591*** | 0.173*** | 0.308*** | 0.399*** | 0.131*** |
| | (0.045) | (0.088) | (0.036) | (0.040) | (0.094) | (0.049) |
| Pct. in oats | 0.762*** | 1.761*** | 0.701*** | 0.593*** | 1.377*** | 0.621*** |
| | (0.099) | (0.195) | (0.085) | (0.070) | (0.144) | (0.076) |
| Pct. in barley | 0.354 | -0.393 | -0.369* | 0.435** | -0.453 | -0.370** |
| | (0.283) | (0.504) | (0.189) | (0.209) | (0.445) | (0.186) |
| Pct. in rye | -0.043 | -1.056 | -0.134 | -2.210*** | -1.767*** | -0.224 |
| | (0.853) | (0.747) | (0.337) | (0.476) | (0.588) | (0.278) |
| N | 1034 | 954 | 954 | 1034 | 941 | 941 |
| $R^2$ | 0.48 | 0.59 | 0.62 | 0.74 | 0.84 | 0.79 |
| RMSE | 0.08 | 0.11 | 0.05 | 0.06 | 0.07 | 0.04 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |
| Min. F-stat | 18.67 | 20.79 | 20.79 | 23.35 | 15.98 | 15.98 |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor in 1930, 1940, and from 1930-1940. All specifications regress tractors per 100 farm acres on contemporaneous crop mixes instrumented with pre-tractor era values. The lowest first stage F-stat is provided. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

## III.D Weighted Regressions

Table III.D.1: Effect of crop mix on tractor diffusion, 1930 and 1940, weighted by farms; OLS

| | Diffusion | | Diff. Change | Diffusion | | Diff. Change |
|---|---|---|---|---|---|---|
| | 1930 | 1940 | 1930-1940 | 1930 | 1940 | 1930-1940 |
| Pct. in corn | 0.035 | 0.621*** | 0.317*** | 0.057 | 0.375*** | 0.152*** |
| | (0.045) | (0.063) | (0.033) | (0.059) | (0.052) | (0.040) |
| Pct. in wheat | 0.873*** | 0.780*** | 0.153*** | 0.626*** | 0.439*** | 0.012 |
| | (0.042) | (0.067) | (0.033) | (0.048) | (0.046) | (0.043) |
| Pct. in oats | 1.112*** | 1.390*** | 0.581*** | 0.895*** | 1.104*** | 0.440*** |
| | (0.065) | (0.091) | (0.056) | (0.073) | (0.071) | (0.058) |
| Pct. in barley | 0.894*** | 1.624*** | 0.626*** | 0.592*** | 1.273*** | 0.613*** |
| | (0.187) | (0.139) | (0.088) | (0.164) | (0.127) | (0.091) |
| Pct. in rye | -1.122*** | -1.582*** | 0.076 | -1.272*** | -1.392*** | -0.137 |
| | (0.317) | (0.337) | (0.155) | (0.261) | (0.258) | (0.144) |
| N | 1034 | 954 | 954 | 1034 | 941 | 941 |
| $R^2$ | 0.70 | 0.75 | 0.69 | 0.80 | 0.91 | 0.76 |
| RMSE | 0.08 | 0.10 | 0.05 | 0.06 | 0.06 | 0.05 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor in 1930, 1940, and from 1930-1940. All specifications regress the fraction of farms with tractors on contemporaneous crop mixes. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Regressions are weighted by farm count. Robust SEs in parentheses.

Table III.D.2: Effect of crop mix on tractor diffusion, 1930 and 1940, weighted by farms; IV

| | Diffusion | | Diff. Change | Diffusion | | Diff. Change |
|---|---|---|---|---|---|---|
| | 1930 | 1940 | 1930-1940 | 1930 | 1940 | 1930-1940 |
| Pct. in corn | -0.014 | 0.360*** | 0.334*** | 0.014 | 0.343*** | 0.290*** |
| | (0.062) | (0.091) | (0.040) | (0.080) | (0.078) | (0.062) |
| Pct. in wheat | 0.549*** | 0.931*** | 0.225*** | 0.359*** | 0.543*** | 0.140** |
| | (0.069) | (0.111) | (0.050) | (0.071) | (0.079) | (0.071) |
| Pct. in oats | 1.197*** | 2.258*** | 0.650*** | 0.863*** | 1.333*** | 0.284*** |
| | (0.094) | (0.152) | (0.085) | (0.090) | (0.118) | (0.101) |
| Pct. in barley | 0.827** | 1.090*** | 0.700*** | 0.520 | 1.187*** | 0.955*** |
| | (0.375) | (0.333) | (0.191) | (0.355) | (0.297) | (0.228) |
| Pct. in rye | -1.499*** | -1.563** | -0.192 | -2.673*** | -2.700*** | -0.718** |
| | (0.580) | (0.673) | (0.365) | (0.512) | (0.532) | (0.323) |
| N | 1034 | 954 | 954 | 1034 | 941 | 941 |
| $R^2$ | 0.68 | 0.73 | 0.68 | 0.79 | 0.90 | 0.74 |
| RMSE | 0.08 | 0.10 | 0.05 | 0.07 | 0.06 | 0.05 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |
| Min. F-stat | 22.01 | 20.86 | 20.86 | 25.09 | 17.66 | 17.66 |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor in 1930, 1940, and from 1930-1940. All specifications regress the fraction of farms with tractors on contemporaneous crop mixes instrumented with pre-tractor era values. The lowest first stage F-stat is provided. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Regressions are weighted by farm count. Robust SEs in parentheses.

Table III.D.3: Effect of crop mix on tractor vintage, 1940, weighted by farms; OLS

| | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 |
|---|---|---|---|---|---|---|
| Pct. in corn | -0.810*** | 0.161*** | 0.649*** | -0.448*** | 0.118*** | 0.330*** |
| | (0.064) | (0.020) | (0.061) | (0.069) | (0.024) | (0.065) |
| Pct. in wheat | -0.284*** | 0.089*** | 0.195*** | -0.029 | 0.038 | -0.009 |
| | (0.052) | (0.021) | (0.045) | (0.068) | (0.024) | (0.064) |
| Pct. in oats | -0.304*** | 0.008 | 0.297*** | -0.590*** | 0.116*** | 0.474*** |
| | (0.104) | (0.033) | (0.099) | (0.100) | (0.036) | (0.089) |
| Pct. in barley | -1.244*** | 0.487*** | 0.757*** | -1.216*** | 0.530*** | 0.685*** |
| | (0.195) | (0.059) | (0.155) | (0.188) | (0.058) | (0.156) |
| Pct. in rye | -0.959*** | -0.256** | 1.215*** | -0.895*** | -0.124 | 1.019*** |
| | (0.293) | (0.104) | (0.233) | (0.301) | (0.102) | (0.242) |
| N | 954 | 954 | 954 | 941 | 941 | 941 |
| $R^2$ | 0.65 | 0.46 | 0.60 | 0.71 | 0.56 | 0.66 |
| RMSE | 0.09 | 0.03 | 0.08 | 0.08 | 0.03 | 0.07 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |

Notes: Table shows the tendency of counties with different crop mixes to own tractors of different vintages in 1940. All specifications regress the frequency of each tractor vintage on contemporaneous crop mixes. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Regressions are weighted by farm count. Robust SEs in parentheses.

Table III.D.4: Effect of crop mix on tractor vintage, 1940, weighted by farms; IV

| | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 | Vintage: Pre-1930 | Vintage: 1931-35 | Vintage: 1936-40 |
|---|---|---|---|---|---|---|
| Pct. in corn | -0.989*** | 0.203*** | 0.787*** | -0.630*** | 0.179*** | 0.451*** |
| | (0.090) | (0.031) | (0.088) | (0.129) | (0.043) | (0.113) |
| Pct. in wheat | -0.311*** | 0.100*** | 0.211*** | -0.245* | 0.127*** | 0.119 |
| | (0.091) | (0.030) | (0.079) | (0.131) | (0.038) | (0.118) |
| Pct. in oats | 0.265 | -0.155*** | -0.109 | 0.067 | -0.174** | 0.107 |
| | (0.174) | (0.059) | (0.163) | (0.182) | (0.068) | (0.152) |
| Pct. in barley | -2.589*** | 0.882*** | 1.707*** | -2.605*** | 1.094*** | 1.510*** |
| | (0.414) | (0.122) | (0.337) | (0.502) | (0.150) | (0.404) |
| Pct. in rye | -0.615 | -0.679*** | 1.294** | 0.416 | -0.607*** | 0.191 |
| | (0.663) | (0.222) | (0.619) | (0.705) | (0.226) | (0.631) |
| N | 954 | 954 | 954 | 941 | 941 | 941 |
| $R^2$ | 0.60 | 0.41 | 0.57 | 0.66 | 0.47 | 0.64 |
| RMSE | 0.09 | 0.04 | 0.08 | 0.09 | 0.03 | 0.07 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |
| Min. F-stat | 20.86 | 20.86 | 20.86 | 17.66 | 17.66 | 17.66 |

Notes: Table shows the tendency of counties with different crop mixes to own tractors of different vintages in 1940. All specifications regress the frequency of each tractor vintage on contemporaneous crop mixes instrumented with pre-tractor era values. The lowest first stage F-stat is provided. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Regressions are weighted by farm count. Robust SEs in parentheses.

## III.E Spatial Standard Errors

Table III.E.1: Comparison to Conley (1999) Standard Errors, 1930 Diffusion; OLS

| | **Coefficient** | **Robust s.e.** | **20-mile Conley s.e.** | **50-mile Conley s.e.** | **100-mile Conley s.e.** |
|---|---|---|---|---|---|
| Pct. in corn | 0.033 | 0.059 | 0.060 | 0.081 | 0.104 |
| Pct. in wheat | 0.695 | 0.043 *** | 0.043 *** | 0.054 *** | 0.062 *** |
| Pct. in oats | 0.995 | 0.082 *** | 0.083 *** | 0.117 *** | 0.156 *** |
| Pct. in barley | 0.795 | 0.152 *** | 0.151 *** | 0.188 *** | 0.229 *** |
| Pct. in rye | -1.270 | 0.247 *** | 0.245 *** | 0.283 *** | 0.315 *** |

Notes: Table compares heteroskedasticity-robust and Conley (1999) standard errors on the main independent variables in a regression of 1930 tractor diffusion on the fraction of farmland in each of the five principal crops and controls. Conley (1999) standard errors allow for spatial correlation in error terms that declines linearly in distance up to a cutoff-point, which is given in the column heading for each set of standard errors. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table III.E.2: Comparison to Conley (1999) Standard Errors, 1940 Diffusion; OLS

| | **Coefficient** | **Robust s.e.** | **20-mile Conley s.e.** | **50-mile Conley s.e.** | **100-mile Conley s.e.** |
|---|---|---|---|---|---|
| Pct. in corn | 0.395 | 0.051 *** | 0.051 *** | 0.067 *** | 0.087 *** |
| Pct. in wheat | 0.477 | 0.046 *** | 0.046 *** | 0.056 *** | 0.068 *** |
| Pct. in oats | 1.129 | 0.070 *** | 0.070 *** | 0.093 *** | 0.116 *** |
| Pct. in barley | 1.141 | 0.135 *** | 0.134 *** | 0.166 *** | 0.206 *** |
| Pct. in rye | -1.166 | 0.247 *** | 0.245 *** | 0.301 *** | 0.323 *** |

Notes: Table compares heteroskedasticity-robust and Conley (1999) standard errors on the main independent variables in a regression of 1940 tractor diffusion on the fraction of farmland in each of the five principal crops and controls. Conley (1999) standard errors allow for spatial correlation in error terms that declines linearly in distance up to a cutoff-point, which is given in the column heading for each set of standard errors. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table III.E.3: Comparison to Conley (1999) Standard Errors, 1930-40 Diffusion Change; OLS

| | **Coefficient** | **Robust s.e.** | **20-mile Conley s.e.** | **50-mile Conley s.e.** | **100-mile Conley s.e.** |
|---|---|---|---|---|---|
| Pct. in corn | 0.164 | 0.043 *** | 0.043 *** | 0.058 *** | 0.079 ** |
| Pct. in wheat | -0.019 | 0.046 | 0.045 | 0.058 | 0.074 |
| Pct. in oats | 0.466 | 0.057 *** | 0.056 *** | 0.072 *** | 0.091 *** |
| Pct. in barley | 0.657 | 0.093 *** | 0.092 *** | 0.112 *** | 0.136 *** |
| Pct. in rye | 0.091 | 0.136 | 0.134 | 0.156 | 0.190 |

Notes: Table compares heteroskedasticity-robust and Conley (1999) standard errors on the main independent variables in a regression of the change in tractor diffusion from 1930 to 1940 on the fraction of farmland in each of the five principal crops and controls. Conley (1999) standard errors allow for spatial correlation in error terms that declines linearly in distance up to a cutoff-point, which is given in the column heading for each set of standard errors. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table III.E.4: Comparison to Conley (1999) Standard Errors, 1930 Diffusion; IV

| | Coefficient | Robust s.e. | | 20-mile Conley s.e. | | 50-mile Conley s.e. | | 100-mile Conley s.e. | |
|---|---|---|---|---|---|---|---|---|---|
| Pct. in corn | 0.056 | 0.076 | | 0.077 | | 0.101 | | 0.124 | |
| Pct. in wheat | 0.483 | 0.062 | *** | 0.063 | *** | 0.085 | *** | 0.105 | *** |
| Pct. in oats | 1.007 | 0.100 | *** | 0.104 | *** | 0.143 | *** | 0.183 | *** |
| Pct. in barley | 0.594 | 0.287 | ** | 0.290 | ** | 0.387 | | 0.469 | |
| Pct. in rye | -2.452 | 0.463 | *** | 0.468 | *** | 0.548 | *** | 0.591 | *** |

Notes: Table compares heteroskedasticity-robust and Conley (1999) standard errors on the main independent variables in a regression of 1930 tractor diffusion on the fraction of farmland in each of the five principal crops, instrumenting with 1910 values, and controls. Conley (1999) standard errors allow for spatial correlation in error terms that declines linearly in distance up to a cutoff-point, which is given in the column heading for each set of standard errors. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table III.E.5: Comparison to Conley (1999) Standard Errors, 1940 Diffusion; IV

| | Coefficient | Robust s.e. | | 20-mile Conley s.e. | | 50-mile Conley s.e. | | 100-mile Conley s.e. | |
|---|---|---|---|---|---|---|---|---|---|
| Pct. in corn | 0.375 | 0.083 | *** | 0.084 | *** | 0.103 | *** | 0.121 | *** |
| Pct. in wheat | 0.568 | 0.087 | *** | 0.087 | *** | 0.106 | *** | 0.127 | *** |
| Pct. in oats | 1.357 | 0.123 | *** | 0.126 | *** | 0.162 | *** | 0.192 | *** |
| Pct. in barley | 0.961 | 0.358 | *** | 0.362 | *** | 0.471 | ** | 0.569 | * |
| Pct. in rye | -1.913 | 0.475 | *** | 0.481 | *** | 0.555 | *** | 0.566 | *** |

Notes: Table compares heteroskedasticity-robust and Conley (1999) standard errors on the main independent variables in a regression of 1940 tractor diffusion on the fraction of farmland in each of the five principal crops, instrumenting with 1910 values, and controls. Conley (1999) standard errors allow for spatial correlation in error terms that declines linearly in distance up to a cutoff-point, which is given in the column heading for each set of standard errors. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively.

Table III.E.6: Comparison to Conley (1999) Standard Errors, 1930-40 Diffusion Change; IV

| | Coefficient | Robust s.e. | | 20-mile Conley s.e. | | 50-mile Conley s.e. | | 100-mile Conley s.e. | |
|---|---|---|---|---|---|---|---|---|---|
| Pct. in corn | 0.397 | 0.074 | *** | 0.075 | *** | 0.094 | *** | 0.122 | *** |
| Pct. in wheat | 0.184 | 0.079 | ** | 0.079 | ** | 0.100 | * | 0.120 | |
| Pct. in oats | 0.182 | 0.118 | | 0.118 | | 0.155 | | 0.193 | |
| Pct. in barley | 1.275 | 0.288 | *** | 0.289 | *** | 0.367 | *** | 0.448 | *** |
| Pct. in rye | -0.437 | 0.282 | | 0.284 | | 0.333 | | 0.358 | |

Notes: Table compares heteroskedasticity-robust and Conley (1999) standard errors on the main independent variables in a regression of the change in tractor diffusion from 1930 to 1940 on the fraction of farmland in each of the five principal crops, instrumenting with 1910 values, and controls. Conley (1999) standard errors allow for spatial correlation in error terms that declines linearly in distance up to a cutoff-point, which is given in the column heading for each set of standard errors. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively.

## III.F Alternative Explanations of Tractor Diffusion

### New Deal Relief and Diffusion in the 1930s

Between the Great Depression, the New Deal, and the Dust Bowl, the 1930s was a tumultuous decade for U.S. agriculture. These trends in 1930s agriculture may have affected tractor adoption and be correlated with regional crop mix, generating bias in the baseline estimates. Here I rule out the possibility that New Deal relief programs explain regional differences in tractor diffusion from 1930 to 1940.

Several New Deal programs provided grant and loan relief to farms across the U.S. beginning in 1933. The principal sources of farm relief were the Agricultural Adjustment Administration (AAA), which paid farmers to take farmland out of production and raise crop prices, and Farm Credit Administration (FCA), which provided emergency farm mortgage debt relief and increased the supply of agricultural credit for all purposes. AAA spending comprised 12.1 percent of all New Deal grants from March 1933 to June 1939, while FCA lending totaled 12 percent of all New Deal loans (Fishback, Kantor, and Wallis 2003). In this section I use the county-level AAA and FCA relief variables of Fishback, Kantor, and Wallis (2003; collected from a 1940 government publication) to isolate any potential confounding effects of New Deal relief.

The "New Deal" hypothesis of Clarke (1991) is that AAA payments and FCA loans landed disproportionately in the hands of corn-growing farmers, aiding in their purchase of tractors. Indeed, though counties concentrated in corn tended to receive AAA relief at the same rate as those concentrated in wheat, corn counties received FCA loans at twice the rate of wheat counties (Table III.F.1). If the delayed diffusion of tractors to the Corn Belt or other regions in the Midwest is primarily due to credit constraints, as Clarke argues, then higher levels of FCA lending should positively affect diffusion.

[Table III.F.1 about here]

Table III.F.3 regresses the county-level change in tractor diffusion from 1930 to 1940 on crop mix, AAA relief spending, and FCA loans. Similar to the alternative specifications of previous tables, column (1) uses contemporaneous crop mix, Column (2) instruments with 1910 crop mix, and column (3) regresses directly on 1910 values. Columns (4) to (6) add controls used in previous tables. The effects of New Deal relief are notably absent from these results: AAA spending and FCA lending add little-to-no additional explanatory power beyond the local crop mix. Controlling for New Deal farm relief does not refute the argument that technical advances were the main force behind tractor diffusion in the 1930s.

[Table III.F.3 about here]

Fishback, Kantor, and Wallis (2003) provide one clue as to why New Deal relief might not have affected tractor purchases: AAA spending and FCA lending favored "large farmers and high-income areas." These farms were more likely to own tractors even in the absence of New Deal relief and less likely to be dependent on government financing for general equipment purchases.

### Effects of the Dust Bowl

The Dust Bowl – a series of dust storms that severely eroded topsoil from Plains farmland in the 1930s – may have also interfered with tractor diffusion in the 1930s. Hornbeck (2012) documents the short- and long-run effects of the Dust Bowl and finds that it had an immediate and persistent effect on agricultural land values. These storms caused widespread crop failure and led to declines in agricultural land value on the order of 30 percent in the most affected counties (Hornbeck 2012). Farmers living in affected counties might have subsequently been unable to afford or borrow against their land to buy a tractor.

Figure 2 of Hornbeck (2012) provides a map of cumulative soil erosion in the 1930s across portions of 12 states in or near the U.S. Plains region. Six of the states and 472 of the counties in this map are in the Midwest sample studied here.[5,6] Hornbeck (2012) digitally traces this map to calculate the fraction of farmland in each county (using 1910 borders) with low, medium, and high erosion.[7] Among the 472 counties in both Hornbeck's (2012) data and the border-adjusted Midwest sample, the average fraction of high-erosion farmland is 0.18; medium-erosion, 0.39; and low-erosion, 0.44. The counties in this subsample were less severely hit by the Dust Bowl than those in Hornbeck's 779-county sample, where the average county has 37, 48, and 15 percent of its farmland in high-, medium-, and low-erosion areas, respectively.

The intersected sample nevertheless includes sufficient variation to distinguish the effects of the Dust Bowl from those of crop mix and technological advances in tractor design. Table III.F.2 shows how the severity of erosion varied with crop concentrations in this sample. Counties concentrated in corn suffered the most severe erosion. Those growing wheat tended to experience moderate erosion; this pattern differs from that of Hornbeck's 779-county sample, over which the percent of farmland in wheat correlates with lower levels of both moderate and severe erosion (Hornbeck 2012, Table 1). Counties growing oats and barley were among the least affected, lying further away from the figurative eye of the storm.

[Table III.F.2 about here]

These patterns suggest that the Dust Bowl cannot explain the relatively larger increases in tractor diffusion from 1930 to 1940 across the Corn Belt, as counties in the heart of this region were among the most severely impacted. Table III.F.4 incorporates the erosion variables into the baseline regressions for the change in diffusion over the decade. Though the IV specification no longer has sufficient explanatory power in the first stage to generate unbiased second-stage estimates, the regressions on 1910 crop mix (Columns 3 and 6) are identified. The results suggest relatively modest, if any, effects of the Dust Bowl on tractor diffusion but continue to highlight the importance of the local crop mix to the technology's spread.

[Table III.F.4 about here]

[5] For all Dust Bowl-related regressions I apply Hornbeck's (2010) border adjustment procedure to reapportion all post-1910 data for my sample to 1910 county borders. This allows me to merge Hornbeck's (2012) erosion variables, which are defined for 1910 county borders, to my data without any additional changes.

[6] The states in the intersected sample are Iowa, Kansas, Minnesota, Nebraska, North Dakota, and South Dakota. Other states shown in Hornbeck's map are Colorado, Montana, New Mexico, Oklahoma, Texas, and Wyoming. Substantial portions of these states suffered severe erosion, especially areas in eastern Colorado, Oklahoma, and central Montana. Most of eastern New Mexico and central Texas also suffered moderate erosion.

[7] "Low erosion" is defined to be less than 25 percent of topsoil lost; "medium erosion" indicates between 25 and 75 percent of topsoil lost, and "high erosion" indicates more than 75 percent of topsoil lost.

Table III.F.1: New Deal Farm Relief (AAA/FCA) by 1930 crop mix

| | (1) AAA spending ($/acre) | (2) FCA loans ($/acre) |
|---|---|---|
| Pct. in corn, 1930 | 8.151*** | 3.052*** |
| | (0.624) | (0.366) |
| Pct. in wheat, 1930 | 8.921*** | 1.519*** |
| | (0.386) | (0.216) |
| Pct. in oats, 1930 | 1.378* | 2.824*** |
| | (0.811) | (0.495) |
| Pct. in barley, 1930 | 9.426*** | 4.794*** |
| | (1.603) | (1.244) |
| Pct. in rye, 1930 | -0.751 | -0.277 |
| | (1.622) | (3.150) |
| Constant | 0.265 | 0.471*** |
| | (0.221) | (0.131) |
| N | 1032 | 1032 |
| $R^2$ | 0.71 | 0.49 |
| RMSE | 0.86 | 0.63 |

Notes: Table shows the rate at which counties with different crop mixes received New Deal relief spending. Column (1) regresses AAA spending per acre of farmland on the fraction of farmland in five principal crops. Column (2) regresses FCA lending per acre on crop percentages. All regressions include state FEs (not shown). *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.F.2: Dust Bowl Soil Erosion (low/med/high) by 1930 crop mix

| | (1) Erosion: Low | (2) Erosion: Med. | (3) Erosion: High |
|---|---|---|---|
| Pct. in corn, 1930 | -1.576*** | 0.360* | 1.216*** |
| | (0.206) | (0.207) | (0.153) |
| Pct. in wheat, 1930 | -0.118 | 0.413*** | -0.294*** |
| | (0.129) | (0.150) | (0.110) |
| Pct. in oats, 1930 | 2.197*** | -0.399 | -1.798*** |
| | (0.411) | (0.401) | (0.321) |
| Pct. in barley, 1930 | 1.738*** | -2.036*** | 0.298 |
| | (0.568) | (0.482) | (0.402) |
| Pct. in rye, 1930 | -0.831 | 2.032 | -1.201* |
| | (1.533) | (1.506) | (0.618) |
| Constant | 0.465*** | 0.389*** | 0.146** |
| | (0.061) | (0.059) | (0.057) |
| N | 471 | 471 | 471 |
| $R^2$ | 0.44 | 0.15 | 0.34 |
| RMSE | 0.29 | 0.31 | 0.24 |

Notes: Table shows the severity of the Dust Bowl across counties with different crop mixes. Columns (1) to (3) regress the fraction of farmland with low, medium, and high cumulative erosion in the 1930s (respectively) on the fraction of farmland in five principal crops. All regressions include state FEs (not shown). *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.F.3: Effect of crop mix on change in tractor diffusion, 1930-40, controlling for New Deal

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Pct. in corn | 0.231*** | 0.352*** | | 0.175*** | 0.466*** | |
| | (0.053) | (0.062) | | (0.048) | (0.093) | |
| Pct. in wheat | 0.070 | 0.246*** | | -0.007 | 0.258*** | |
| | (0.044) | (0.068) | | (0.046) | (0.096) | |
| Pct. in oats | 0.630*** | 0.632*** | | 0.458*** | 0.114 | |
| | (0.052) | (0.096) | | (0.059) | (0.126) | |
| Pct. in barley | 0.691*** | 0.928*** | | 0.657*** | 1.385*** | |
| | (0.089) | (0.209) | | (0.093) | (0.297) | |
| Pct. in rye | 0.336** | 0.007 | | 0.083 | -0.511* | |
| | (0.140) | (0.328) | | (0.138) | (0.294) | |
| Pct. in corn, 1910 | | | 0.241*** | | | 0.356*** |
| | | | (0.039) | | | (0.050) |
| Pct. in wheat, 1910 | | | 0.142*** | | | 0.120*** |
| | | | (0.037) | | | (0.041) |
| Pct. in oats, 1910 | | | 0.581*** | | | 0.293*** |
| | | | (0.045) | | | (0.053) |
| Pct. in barley, 1910 | | | 0.665*** | | | 0.676*** |
| | | | (0.114) | | | (0.113) |
| Pct. in rye, 1910 | | | -0.088 | | | -0.481** |
| | | | (0.275) | | | (0.236) |
| AAA spending ($/acre) | 0.007** | -0.001 | 0.003 | -0.002 | -0.011*** | -0.005* |
| | (0.003) | (0.003) | (0.003) | (0.003) | (0.004) | (0.003) |
| FCA loans ($/acre) | -0.002 | -0.002 | 0.002 | 0.002 | 0.004 | 0.004 |
| | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) | (0.003) |
| Constant | 0.075*** | 0.062*** | 0.048*** | -0.681* | -1.182** | -0.821** |
| | (0.009) | (0.011) | (0.009) | (0.351) | (0.546) | (0.369) |
| N | 953 | 953 | 1033 | 940 | 940 | 1015 |
| $R^2$ | 0.67 | 0.65 | 0.65 | 0.73 | 0.68 | 0.70 |
| RMSE | 0.05 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |
| Min. F-stat | | 19.71 | | | 16.29 | |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor from 1930-1940. Columns (1) and (4) regress the fraction of farms with tractors on contemporaneous crop mixes. Columns (2) and (5) instrument with pre-tractor era crop mixes. Columns (3) and (6) regress on lagged crop percentages. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

Table III.F.4: Effect of crop mix on change in tractor diffusion, 1930-40, controlling for Dust Bowl

| | (1) | (2) | (3) | (4) | (5) | (6) |
|---|---|---|---|---|---|---|
| Pct. in corn | 0.056 | 0.170** | | 0.065 | 0.519** | |
| | (0.047) | (0.085) | | (0.071) | (0.227) | |
| Pct. in wheat | 0.026 | 0.165* | | -0.114** | 0.277* | |
| | (0.043) | (0.090) | | (0.056) | (0.148) | |
| Pct. in oats | 0.794*** | 0.518** | | 0.673*** | 0.036 | |
| | (0.079) | (0.213) | | (0.087) | (0.200) | |
| Pct. in barley | 1.134*** | 1.703*** | | 0.737*** | 1.296** | |
| | (0.115) | (0.436) | | (0.137) | (0.612) | |
| Pct. in rye | 0.305** | -0.515 | | 0.014 | -1.507** | |
| | (0.153) | (0.604) | | (0.178) | (0.742) | |
| Pct. in corn, 1910 | | | 0.258*** | | | 0.520*** |
| | | | (0.054) | | | (0.082) |
| Pct. in wheat, 1910 | | | 0.108** | | | 0.037 |
| | | | (0.045) | | | (0.057) |
| Pct. in oats, 1910 | | | 0.607*** | | | 0.320*** |
| | | | (0.077) | | | (0.088) |
| Pct. in barley, 1910 | | | 0.706*** | | | 0.501*** |
| | | | (0.130) | | | (0.133) |
| Pct. in rye, 1910 | | | -1.400* | | | -1.840*** |
| | | | (0.714) | | | (0.610) |
| Erosion: Med. | -0.009 | -0.018 | -0.032*** | -0.007 | -0.017 | -0.028** |
| | (0.011) | (0.012) | (0.012) | (0.012) | (0.015) | (0.012) |
| Erosion: High | -0.003 | -0.023 | -0.018 | -0.019 | -0.013 | -0.016 |
| | (0.017) | (0.022) | (0.016) | (0.018) | (0.026) | (0.017) |
| Constant | 0.123*** | 0.132*** | 0.112*** | 1.633 | -3.950* | -1.334 |
| | (0.015) | (0.028) | (0.013) | (1.648) | (2.326) | (1.897) |
| N | 446 | 446 | 472 | 441 | 441 | 467 |
| $R^2$ | 0.60 | 0.54 | 0.55 | 0.66 | 0.51 | 0.60 |
| RMSE | 0.06 | 0.07 | 0.07 | 0.06 | 0.07 | 0.07 |
| State FEs? | Yes | Yes | Yes | Yes | Yes | Yes |
| Controls? | No | No | No | Yes | Yes | Yes |
| Min. F-stat | | 9.64 | | | 4.65 | |

Notes: Table shows the tendency of counties with different crop mixes to adopt the farm tractor from 1930-1940. Columns (1) and (4) regress the fraction of farms with tractors on contemporaneous crop mixes. Columns (2) and (5) instrument with pre-tractor era crop mixes. Columns (3) and (6) regress on lagged crop percentages. Columns (4)-(6) add controls. *, **, *** represent significance at the 0.1, 0.05, and 0.01 levels, respectively. Robust SEs in parentheses.

## III.G Additional Maps

Figure III.G.1: Percent of farmland in corn, 1910

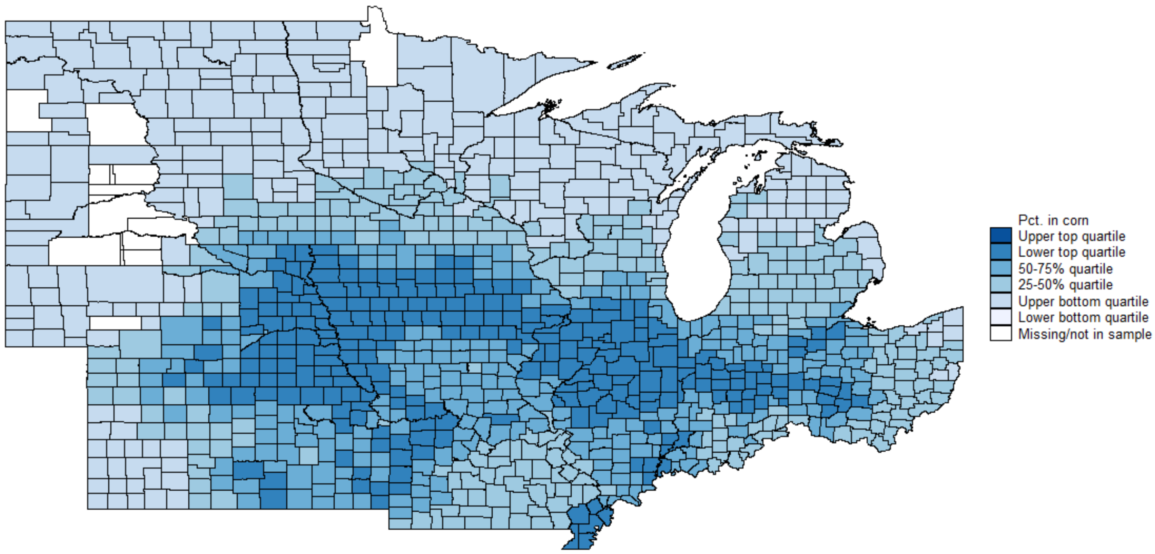



Figure III.G.2: Percent of farmland in corn, 1920

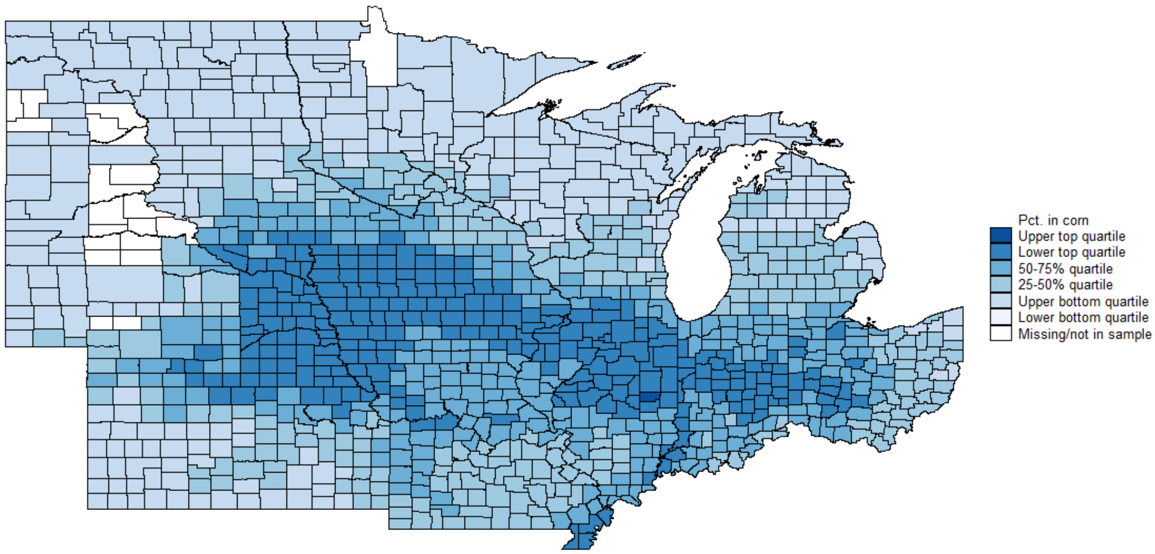


Figure III.G.3: Percent of farmland in corn, 1930

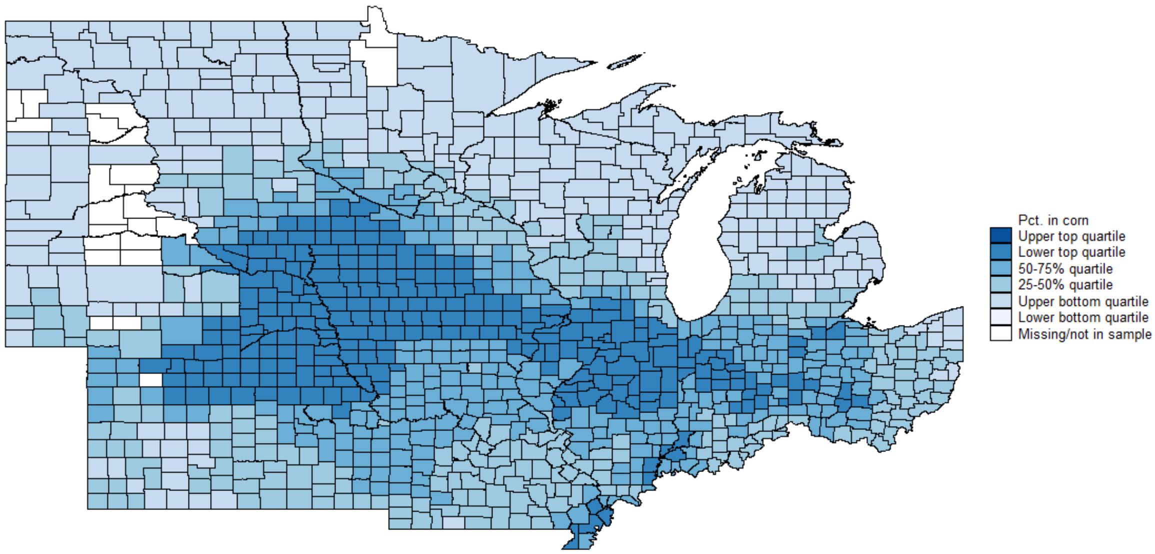



Figure III.G.4: Percent of farmland in corn, 1940

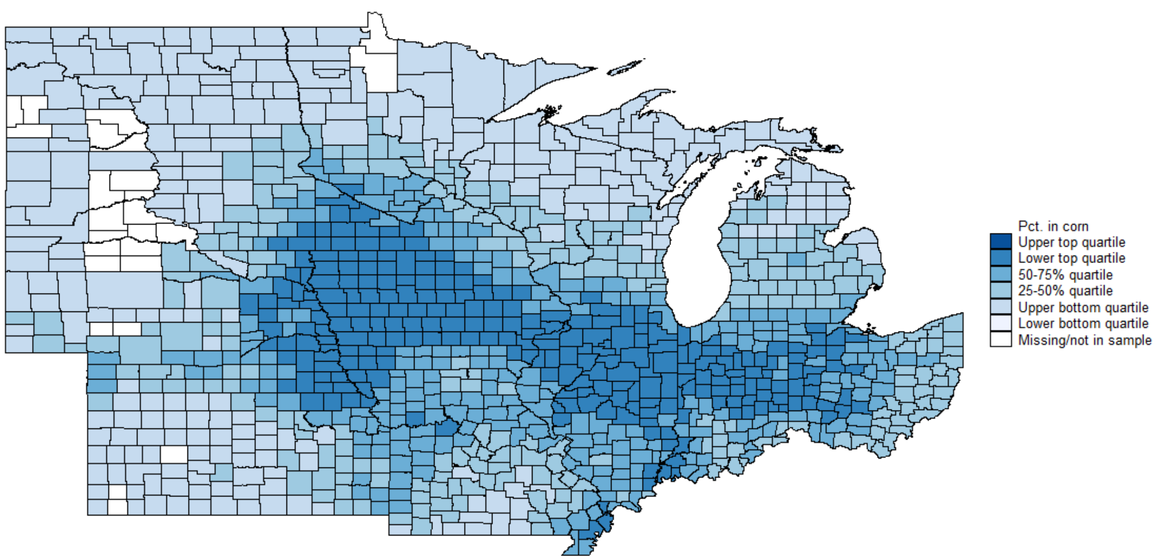


Figure III.G.5: Percent of farmland in wheat, 1910

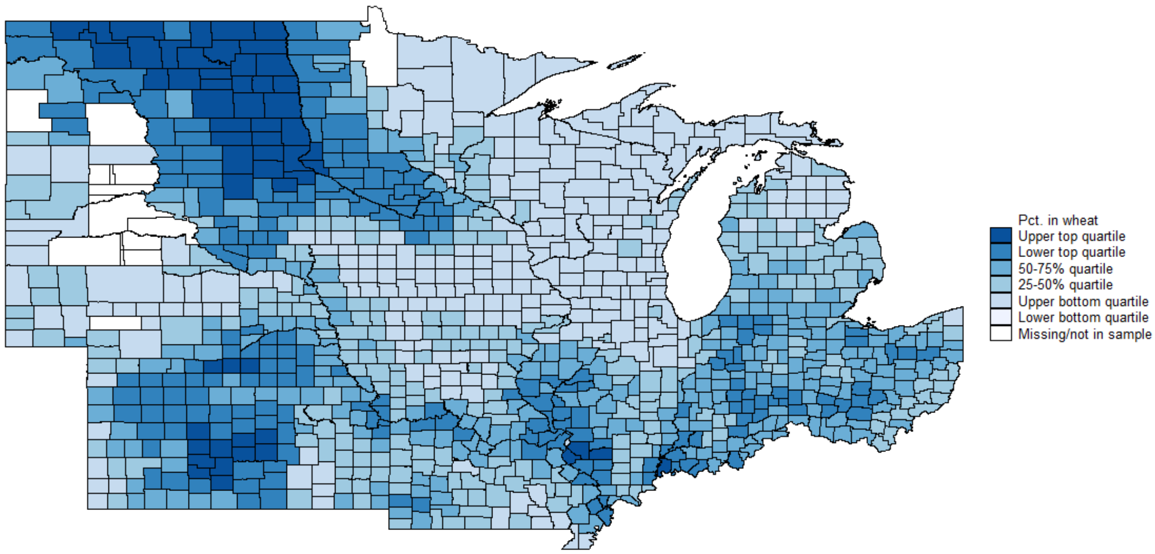



Figure III.G.6: Percent of farmland in wheat, 1920

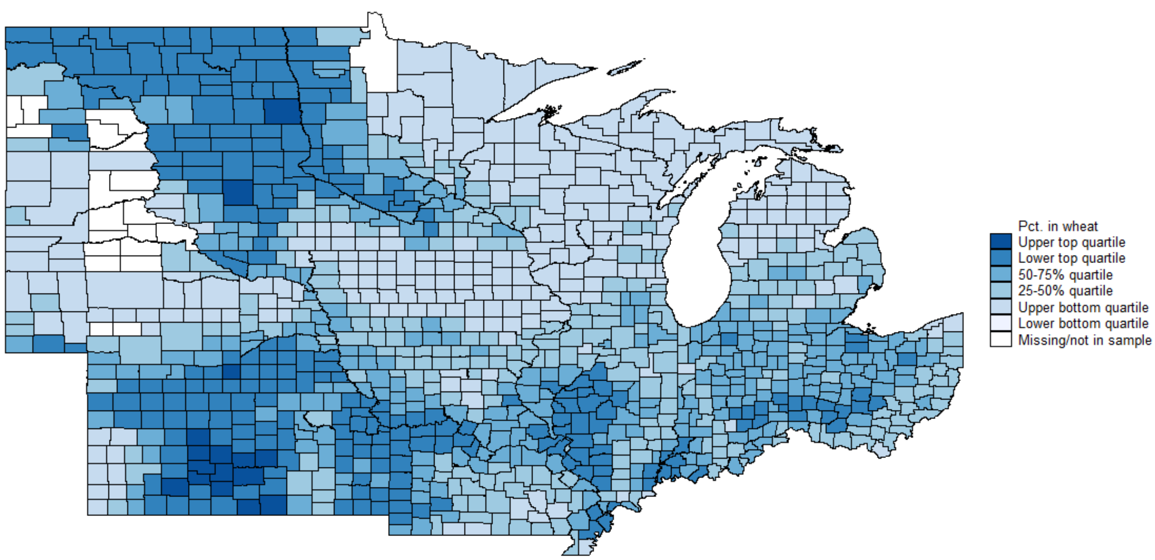


Figure III.G.7: Percent of farmland in wheat, 1930

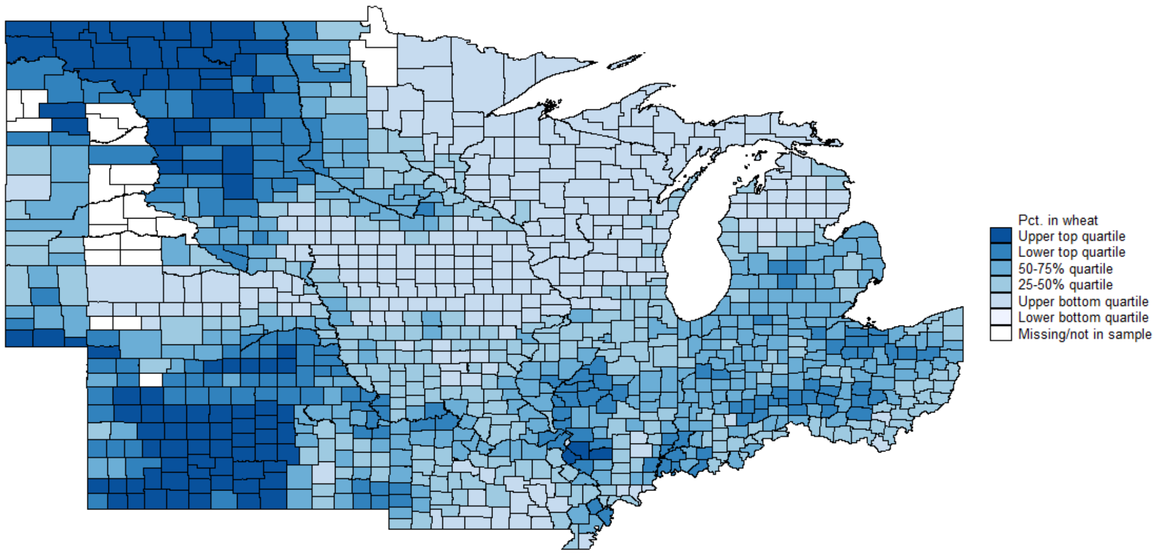



Figure III.G.8: Percent of farmland in wheat, 1940

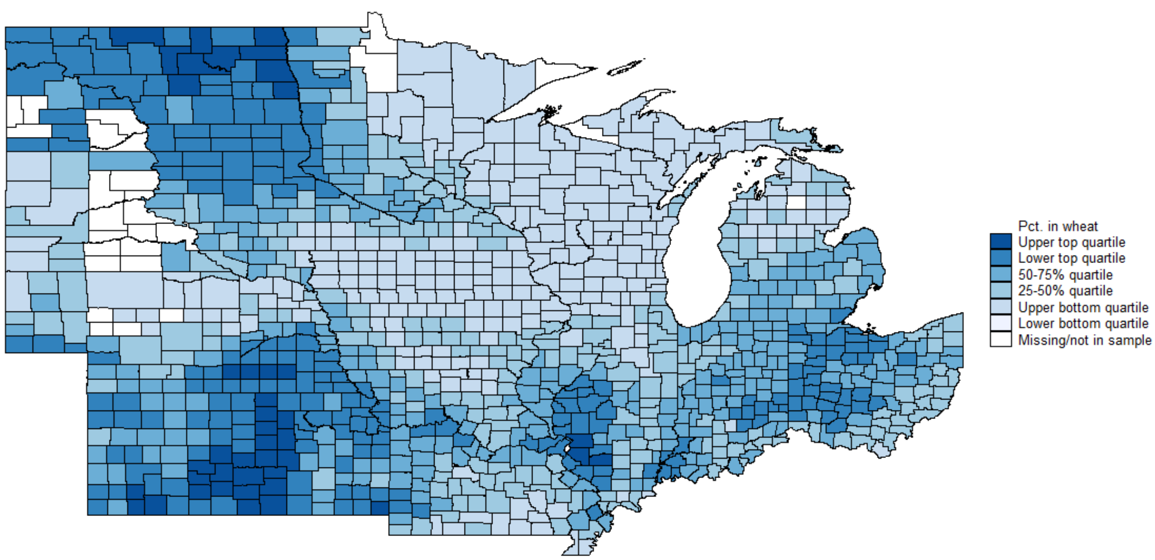


Figure III.G.9: Percent of farmland in oats, 1910

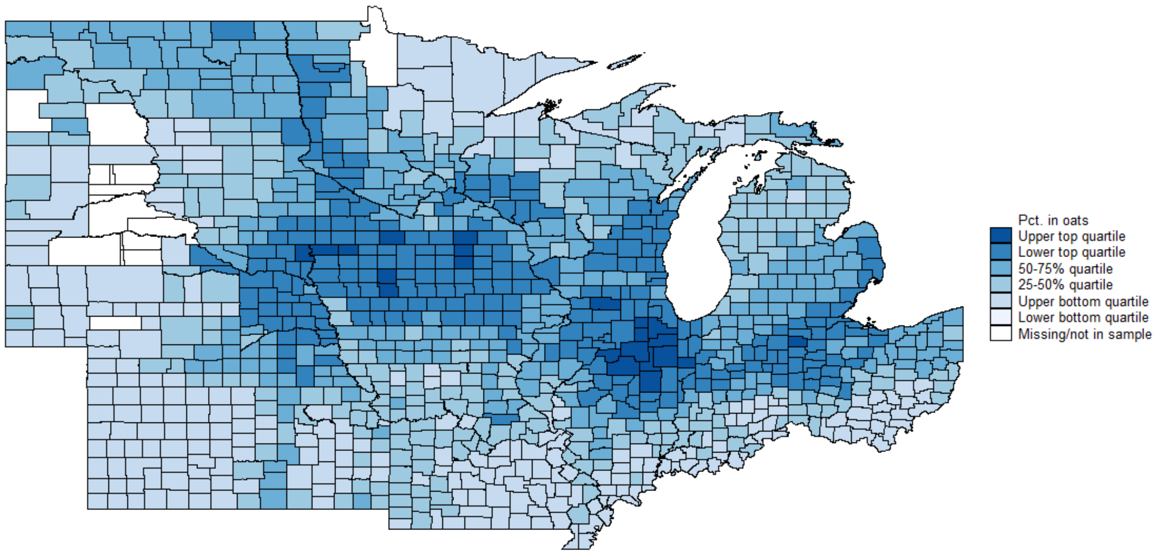



Figure III.G.10: Percent of farmland in oats, 1920

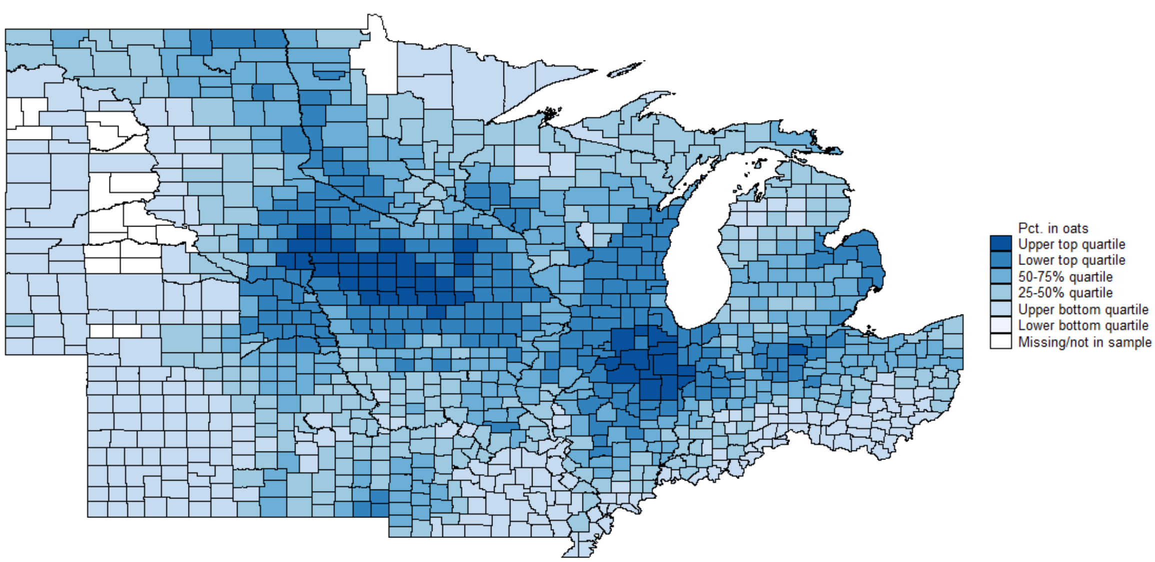


Figure III.G.11: Percent of farmland in oats, 1930

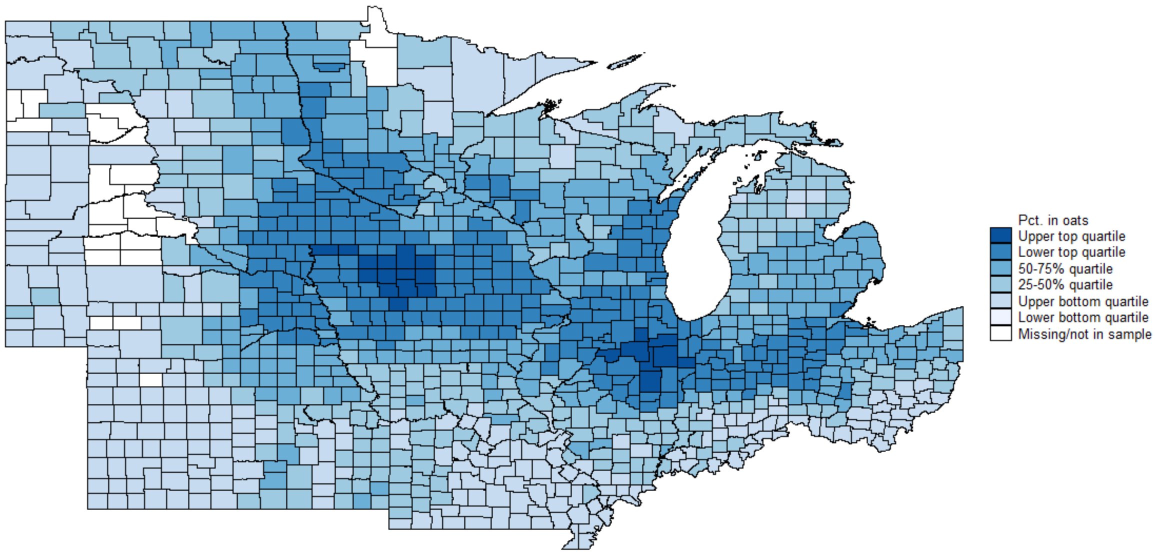



Figure III.G.12: Percent of farmland in oats, 1940

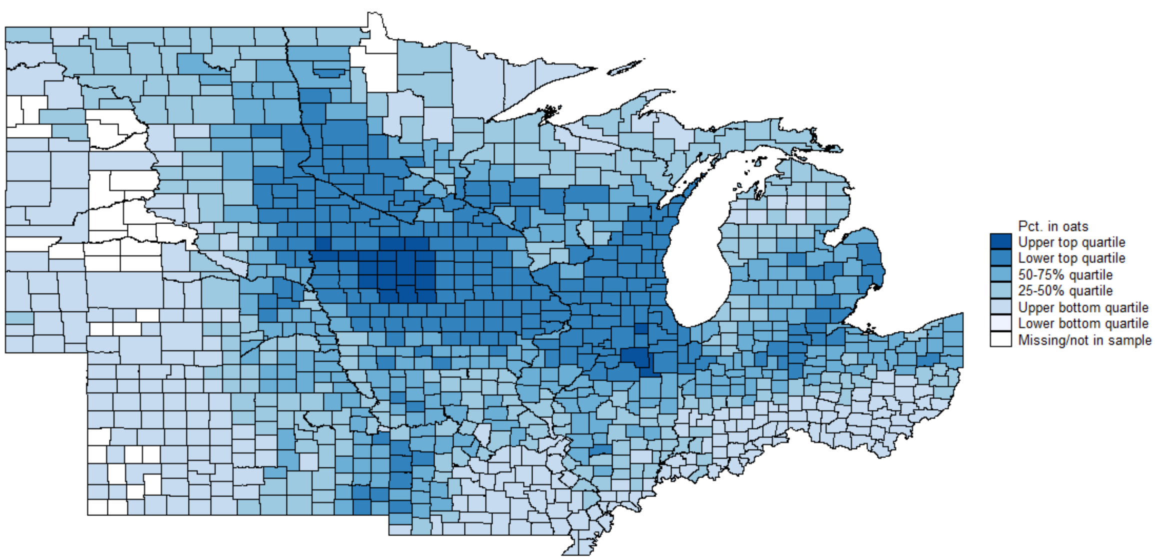


Figure III.G.13: Percent of farmland in barley, 1910

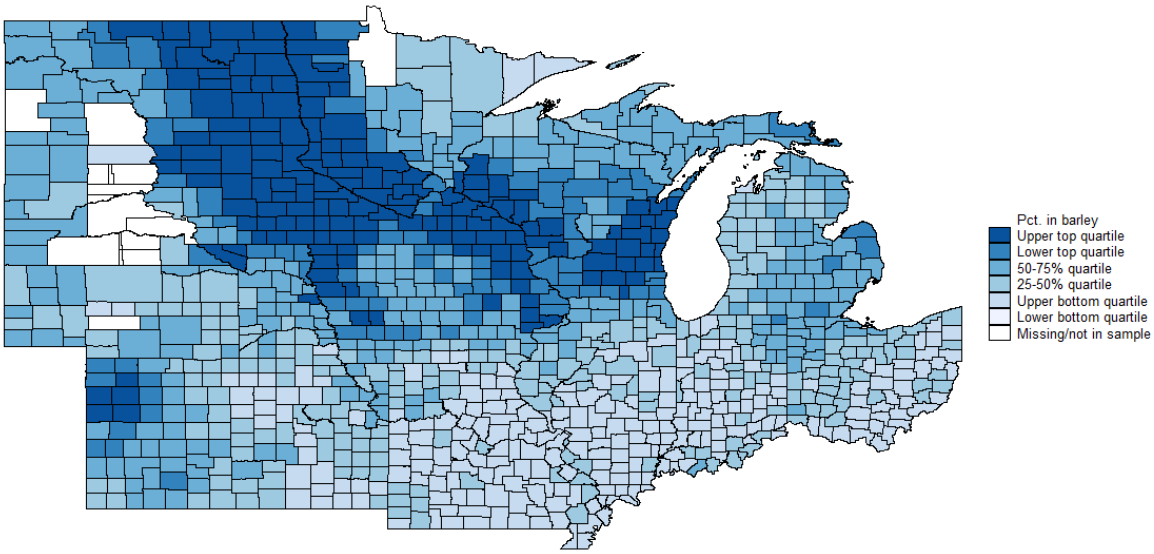



Figure III.G.14: Percent of farmland in barley, 1920

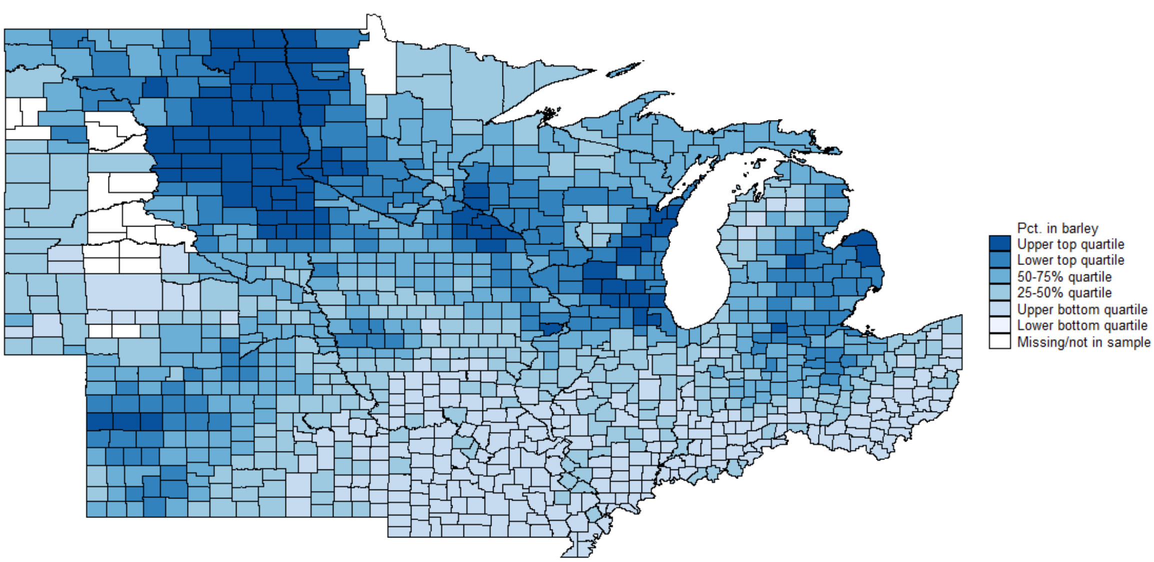


Figure III.G.15: Percent of farmland in barley, 1930

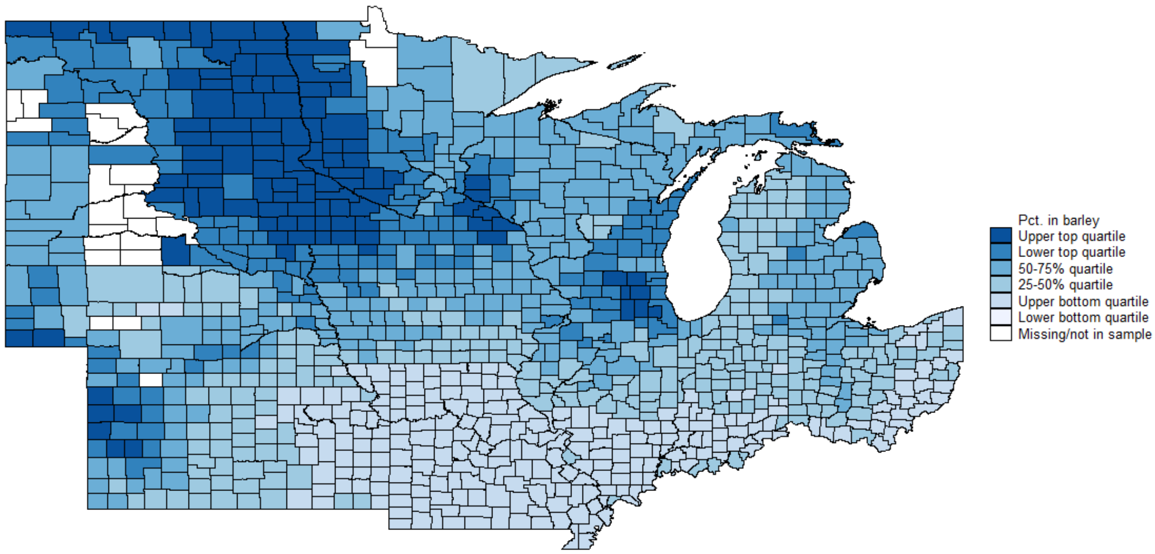



Figure III.G.16: Percent of farmland in barley, 1940

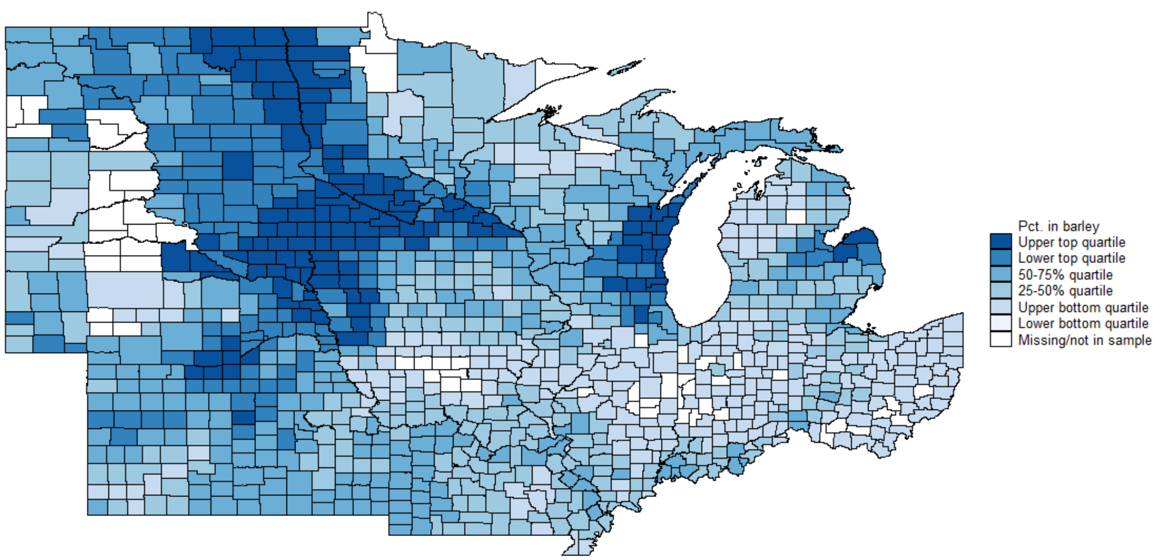


Figure III.G.17: Percent of farmland in rye, 1910

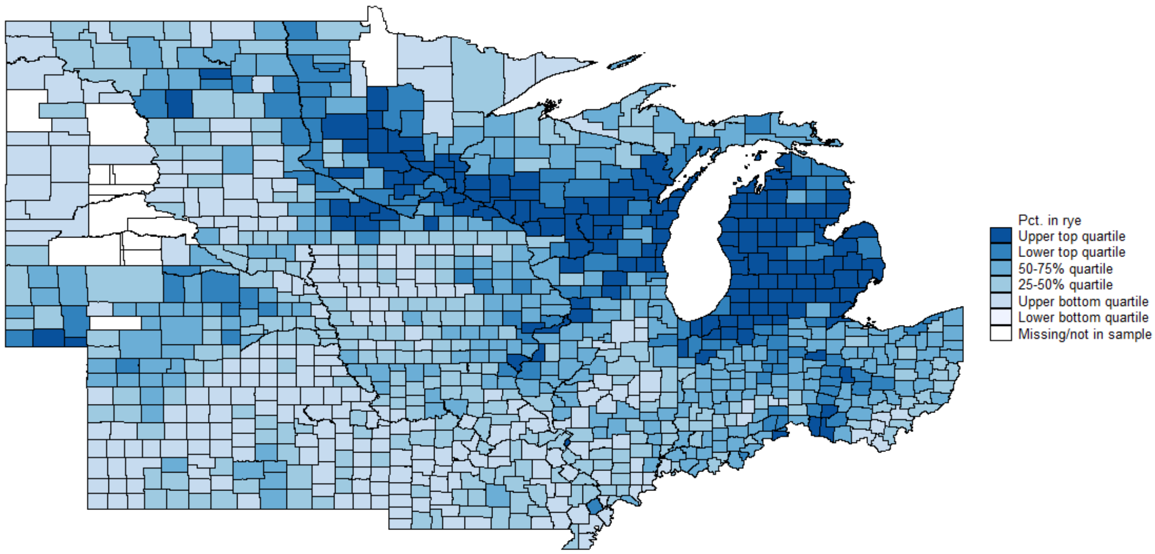



Figure III.G.18: Percent of farmland in rye, 1920

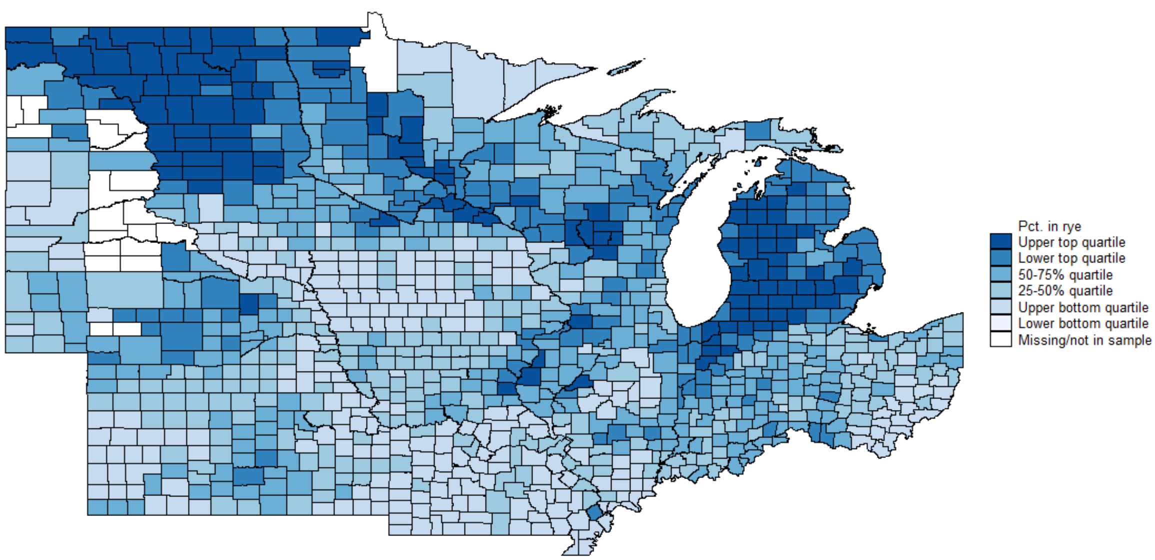


Figure III.G.19: Percent of farmland in rye, 1930

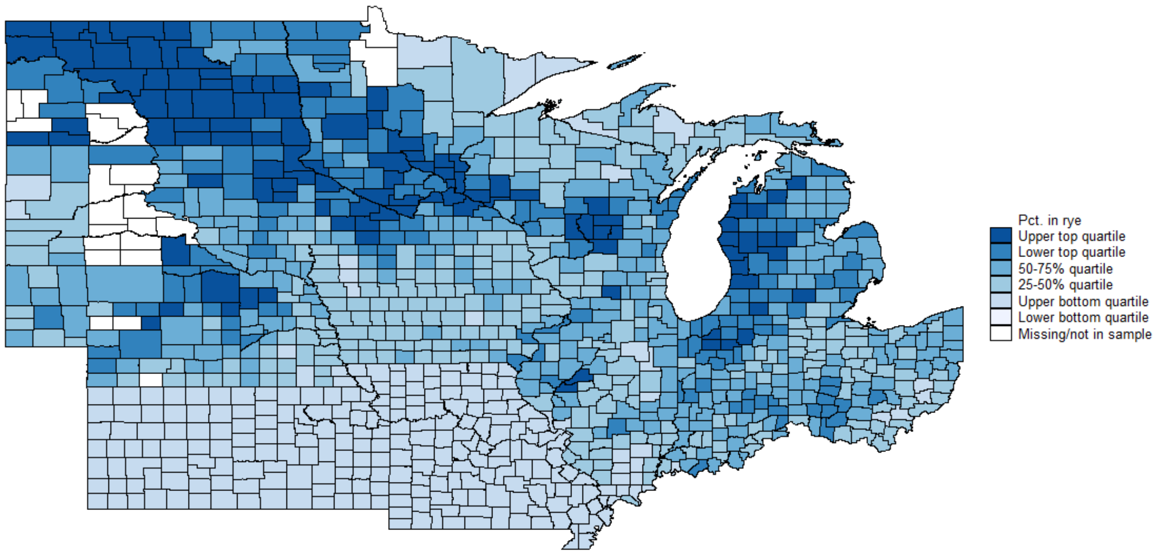



Figure III.G.20: Percent of farmland in rye, 1940

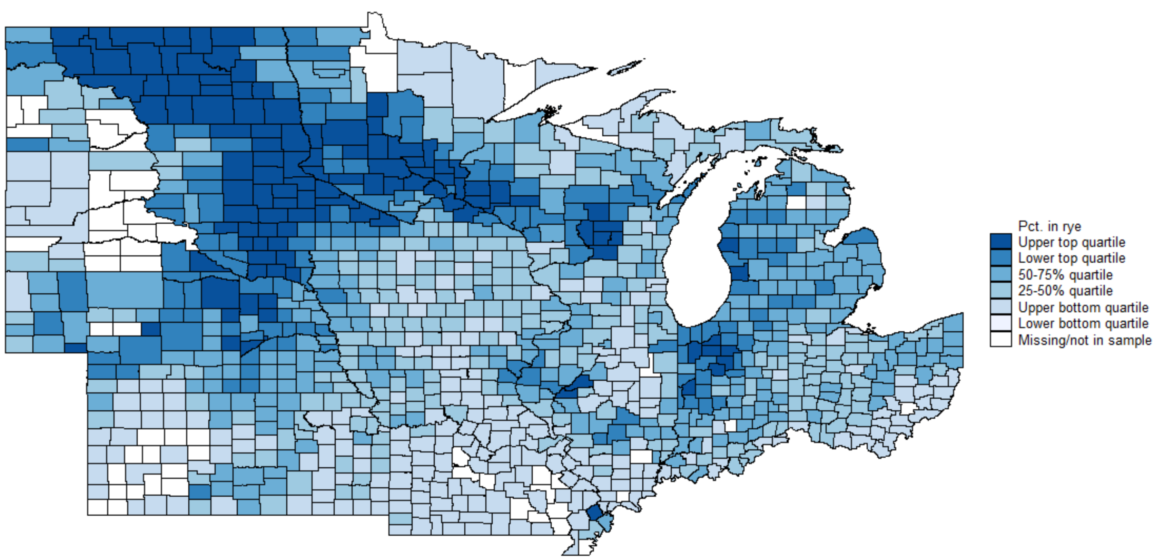


Figure III.G.21: Percent of farmland in hay, 1910

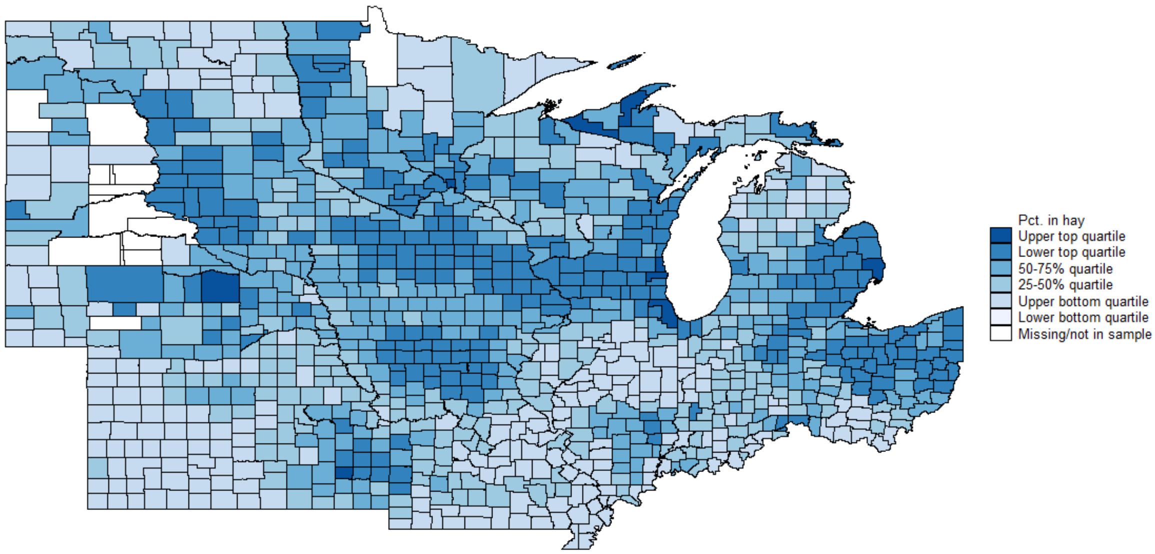



Figure III.G.22: Percent of farmland in hay, 1920

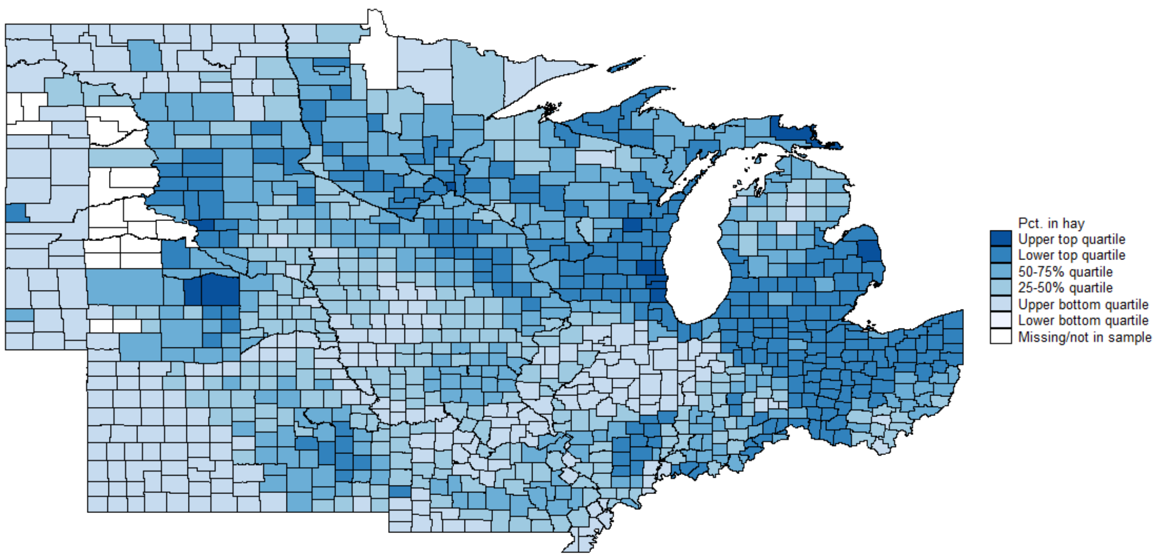


Figure III.G.23: Percent of farmland in hay, 1930

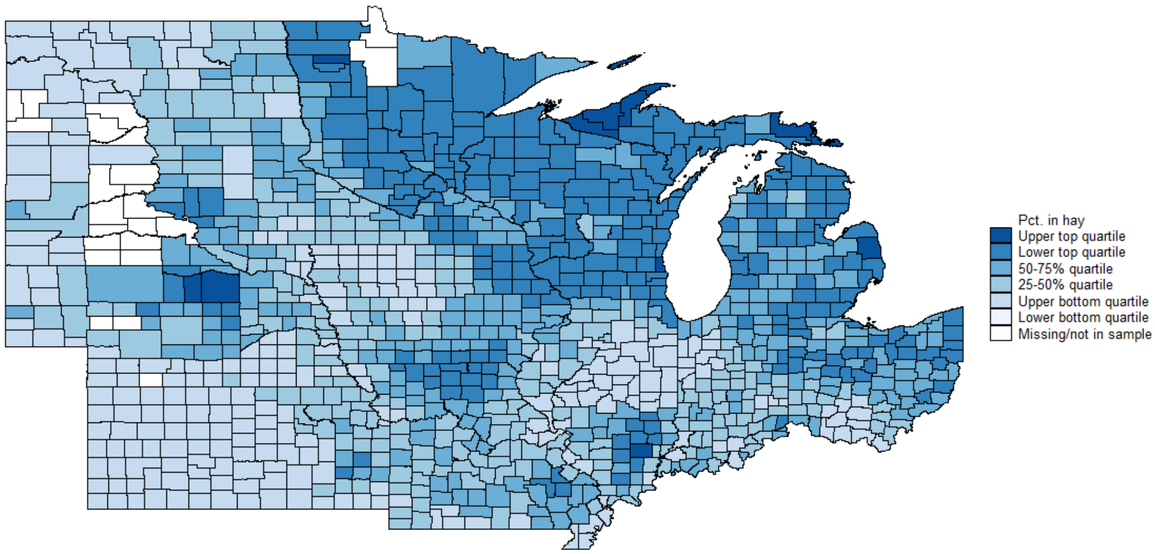



Figure III.G.24: Percent of farmland in hay, 1940