

The Impact of Interpersonal Relationships and Incentive Structures on the Performance of Actors in Informal Supply Chains

by

Olumurejiwa Adedapo Fatunde

A.B., Harvard University (2012)

M.Sc., London School of Economics and Political Science (2013)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Operations Management & Decision Sciences
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author
Department of Civil and Environmental Engineering
May 13, 2022

Certified by.....
Joann F. de Zegher
Maurice F. Strong Career Development Professor
Assistant Professor of Operations Management
MIT Sloan School of Management
Thesis Supervisor

Certified by.....
Yossi Sheffi
Director, MIT Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, MIT Department of Civil and Environmental Engineering
Professor, MIT Institute of Data Science and Society

Accepted by
Colette L. Heald
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee

The Impact of Interpersonal Relationships and Incentive Structures on the Performance of Actors in Informal Supply Chains

by

Olumurejiwa Adedapo Fatunde

Submitted to the Department of Civil and Environmental Engineering
on May 13, 2022, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Management & Decision Sciences

Abstract

This dissertation examines operational challenges faced by intermediaries in informal supply chains, in which the relational and structural constraints present in traditional supply chains are relaxed.

This research is organized into three papers, the first of which (Chapter 2) explores business relationships in the context of emerging market retail supply chains. Attempts to distribute durable, life-improving goods to customers at the Base of the Pyramid (BoP) have struggled to succeed at scale. One potential explanation is poor relationship management with informal retailers, which are embedded within communities. By analyzing data from a distributor selling to 331 formal and 493 informal retailers in India, we demonstrate that informal retailers recover more slowly than formal retailers after a sales agent reallocation. This indicates that disruptions to social/business relationships are particularly harmful when selling to retailers in informal markets.

The second and third papers (Chapters 3 and 4) explore incentive design for distributed-task platforms. We use as a case study a supply chain for medical knowledge, featuring “informal” suppliers without formal contracts. Using data on 5,418 crowdsourcing contests for medical diagnosis, we examine how evaluation metrics (Chapter 3) and prize allocation mechanisms (Chapter 4) shape participants’ decisions and performance. Chapter 3 assesses the impact of evaluating participants using the longest “streak” of correct answers, rather than an accuracy-based metric. Streak evaluation increases volume of quality responses and speed of achieving consensus, largely through increased engagement. These findings are relevant in settings where streak-based-rewards are used to boost motivation; we find that they also boost performance.

Chapter 4 studies how changing the source of prize-related uncertainty from the *probability of winning* to the *amount at stake* affects decision-making. We evaluate the impact of running a *pool* contest (in which participants who meet a performance threshold share prizes evenly) instead of a *rank-order* contest (in which prize distribution is determined exogenously and announced upfront). In pool contests, accuracy increases for average participants but decreases for top performers, suggesting that participants modify engagement levels in re-

sponse to performance thresholds. This suggests that pool contests with carefully-selected thresholds can incentivize effort from participants with certain performance profiles.

Thesis Supervisor: Joann F. de Zegher
Title: Maurice F. Strong Career Development Professor
Assistant Professor of Operations Management
MIT Sloan School of Management

Doctoral Committee Chair: Stephen Graves
Title: Abraham J. Siegel Professor of Management
Professor of Operations Management
MIT Sloan School of Management
Professor, MIT Department of Mechanical Engineering

Doctoral Committee Member: Daniel Frey
Title: MIT D-Lab Faculty Research Director
Professor, MIT Department of Mechanical Engineering

Doctoral Committee Member: Jarrod Goentzel
Title: Director, MIT Humanitarian Supply Chain Lab
Principal Research Scientist

Doctoral Committee Member: Yossi Sheffi
Title: Director, MIT Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, MIT Department of Civil and Environmental Engineering
Professor, MIT Institute of Data Science and Society

Acknowledgments

I would like to thank my advisor, Professor Joann de Zegher, for her tireless encouragement and support throughout this process.

I would also like to thank the Center for Transportation and Logistics (CTL) for serving as my home throughout my time at MIT. Thanks in particular to Dr. Jarrod Goentzel and Professor Yossi Sheffi for supporting my development as a researcher. I am also grateful to the chair of my doctoral committee, Professor Stephen Graves, and to Professor Daniel Frey, who also served as a committee member.

Thank you to the fellow PhD students (at CTL, MIT, and beyond) who have served as a support system over the past five years –it’s been a pleasure sharing this journey with you all.

Thank you to my many mentors and collaborators (particularly Professor Gonzalo Romero and Professor Andre Calmon) for challenging me and helping to accelerate my growth as a scholar, and to each of my partner organizations for entrusting me with your challenges.

Finally, I thank my family and friends for your understanding and support over the past several years.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

1	Introduction	17
1.1	Informal retail supply chains	19
1.2	Crowdsourcing: widely distributed digital supply chains	19
1.3	Dissertation overview	20
2	The value of long-term relationships when selling to informal retailers - Evidence from India	23
2.1	Abstract	23
2.2	Introduction	24
2.3	Data and Key Definitions	27
2.3.1	Retailer order data	29
2.3.2	Retailer tax registration	31
2.3.3	Sales executive employment	32
2.3.4	Retailer-level characteristics	32
2.4	Methodology	33
2.4.1	Within-unit matching	34
2.4.2	Between-unit matching/weighting	37
2.5	Results	44
2.5.1	Within-unit matching	44
2.5.2	Between-unit matching/weighting	46
2.6	Discussion & Concluding Remarks	47
2.6.1	Alternate Time Units	51

3	The impact of streak-based performance evaluation in crowdsourcing con-	
	tests for skilled microtasks	53
3.1	Abstract	53
3.2	Introduction and Literature	54
3.3	Data and Key Definitions	57
3.3.1	Contest-level data	58
3.3.2	Topic-level data	58
3.3.3	Problem-level data	59
3.3.4	Key decisions	60
3.3.5	Description of contest evaluation metrics	61
3.3.6	Treatment variable	65
3.3.7	Outcome variables of interest	65
3.4	Methodology	68
3.4.1	Causal effect of interest	68
3.4.2	Matching	69
3.4.3	Effect estimation	72
3.4.4	Heterogeneous Treatment Effects	76
3.5	Results	78
3.5.1	Primary outcome metrics	78
3.5.2	Secondary outcome metrics	79
	Participation	79
	Engagement	80
	Performance	82
	Speed	82
3.5.3	Effect heterogeneity	83
3.6	Discussion & Concluding Remarks	87
3.7	Appendix	91
3.7.1	Explanatory variables	91
3.7.2	Alternate matching results	92

4	Competitive uncertainty vs prize value uncertainty: the impact of prize endogeneity on participant behavior in crowdsourcing contests	95
4.1	Abstract	95
4.2	Introduction	96
4.2.1	Past literature	98
4.2.2	Research setting	100
4.3	Data and Key Definitions	101
4.3.1	Description of alternative prize structures	103
4.3.2	Treatment variable	107
4.3.3	Explanatory variables	107
4.3.4	Outcome variables of interest	109
4.4	Methodology	111
4.4.1	Hypotheses	111
	Expected earnings for a rank-order contest	111
	Expected earnings for a pool contest	112
	Predictions	115
4.4.2	Causal effect of interest	116
4.4.3	Matching	117
4.4.4	Effect estimation	120
4.4.5	Heterogeneous Treatment Effects	122
4.5	Results	124
4.5.1	Primary outcome metrics	124
4.5.2	Secondary outcome metrics	124
	Participation	124
	Engagement	127
	Performance	128
	Speed	129
4.5.3	Effect heterogeneity	130
4.6	Discussion & Concluding Remarks	132
4.7	Appendix	136

4.7.1	Alternate matching results	136
5	Conclusion	139
5.1	Managerial Insights	139
5.2	Future Research Directions	140

List of Figures

1	Retailer locations	28
2	Distribution of number of transitions across retailers.	30
3	Time at which treated retailers received treatment, and potential controls using within-unit matching.	35
4	Treated retailer and control retailers for between-unit k-nearest-neighbor matching estimator with all potential control units (left panel) and with matched control units only (right panel).	38
5	Balance on key covariates by number of permitted matches based on \overline{B}_{j,t_0-1}^m	43
6	ATT based on the within-unit estimator.	45
7	ATT and 95% confidence intervals for formal and informal retailers using the between-unit estimator.	47
8	Within-unit ATT with alternate time unit (months)	51
1	Example of information available to participants on the outer app screen.	61
2	Example information available to participants on scoring guidelines.	61
3	Median contest scores and length of longest success streak for contests using accuracy-based vs. streak-based performance metric	63
4	Discretized version of minimum prize value	70
5	Simulated expected value of treatment effect for primary outcomes	79
6	Simulated expected value of treatment effect for participation outcomes	80
7	Simulated expected value of treatment effect for engagement outcomes	81
8	Simulated expected value of treatment effect for performance outcomes	82
9	Simulated expected value of treatment effect for speed outcomes	83

10	Conditional ATT based on minimum prize value	85
11	Conditional ATT based on number of prizes available	86
12	Conditional ATT based on total prize basket value	87
1	a) Representative scores for each prize structure, b) Representative scores for each prize structure with pool contests split by performance threshold, and c) Representative earnings for each prize structure	106
2	Example of information available to participants on the outer screen for a) rank-order and b) pool contests	107
3	Example of information available to participants about the prize allocation within a a) rank-order or b) pool contest	109
4	Level of activity over time	109
5	Simulated expected value of treatment effect for primary outcomes	124
6	Simulated expected value of treatment effect	125
7	Average number of pool contests completed a) across all survey respondents and b) across survey respondents with at least one pool contest	126
8	Simulated expected value of treatment effect	128
9	Simulated expected value of treatment effect	129
10	Simulated expected value of treatment effect	130
11	Conditional ATT based on minimum prize value	131
12	Conditional ATT based on minimum prize value	132

List of Tables

1	Informal supply chains as a relaxation of traditional supply chains	18
1	Product categories sold by the distributor (April 2016-December 2019) . . .	29
2	Breakdown of retailers in our sample	31
3	Comparison of survey results for 127 “informal” and 125 “formal” retailers. .	32
4	Covariate balance before and after k-nearest-neighbor matching using values $k = 5$ for formal retailers and $k = 7$ for informal retailers.	43
5	Number of retailers included after matching for formal and informal retailers	44
1	Exclusion criteria for contests on the platform	57
2	Descriptive statistics for explanatory variables ($n = 1488$)	59
3	Contest-level outcome variables.	66
4	Relationship between secondary outcomes and primary outcomes.	67
5	Summary of matching process data sample for outcomes using the complete dataset	71
6	Pre/Post-matching balance summary for contest covariates ($n = 1,488$) . . .	72
7	Statistical distributions used to model each outcome variable	73
8	Subgroups for heterogeneity analysis	78
9	Coefficients, standard errors and p -values for the treatment variable for pri- mary outcomes	79
10	Coefficients, standard errors and p -values for the treatment variable for par- ticipation outcomes	79

11	Coefficients, standard errors and p -values for the treatment variable for engagement outcomes	81
12	Coefficients, standard errors and p -values for the treatment variable for performance outcomes	82
13	Coefficients, standard errors and p -values for the treatment variable for speed outcomes	83
14	Summary of effect heterogeneity.	84
15	Summary of results.	88
16	Available contest-level variables of interest	91
17	Summary of matching process data sample for outcomes with $n = 1477$ (1476)	92
18	Pre/Post-matching balance summary for contest-level covariates ($n = 1477$ (1476))	93
19	Summary of matching process data sample for outcomes with $n = 1117$. . .	94
20	Pre/Post-matching balance summary for contest-level covariates ($n = 1177$) .	94
1	Exclusion criteria for contests on the platform	103
2	Available contest-level variables of interest	108
3	Descriptive statistics for explanatory variables	108
4	Outcome variables of interest	110
5	Sources of uncertainty for participant j in contest k	114
6	Summary of matching process data sample	119
7	Pre/Post-matching balance summary for contest-level covariates	119
8	Statistical distributions used to model each outcome variable	120
9	Coefficients, standard errors and p -values for the treatment variable for primary outcomes	124
10	Raw regression coefficients for "participation" outcomes	125
11	Raw regression coefficients for "engagement" outcomes	128
12	Raw regression coefficients for "performance" outcomes	129
13	Raw regression coefficients for "speed" outcomes	130
14	Summary of effect heterogeneity.	131
15	Summary of results.	133

16	Summary of matching process data sample for outcomes with $n = 1291$ (1290)	136
17	Pre/Post-matching balance summary for contest-level covariates ($n = 1291$ (1290))	137
18	Summary of matching process data sample for outcomes with $n = 1276$. . .	137
19	Pre/Post-matching balance summary for contest-level covariates ($n = 1276$) .	137
20	Summary of matching process data sample for outcomes with $n = 881$	138
21	Pre/Post-matching balance summary for contest-level covariates ($n = 881$) .	138

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Introduction

This dissertation explores the influence of relationships and incentives on the behavior of actors in informal supply chains.

The concept of “informality” within supply chains has appeared in academic literature in the context of smallholder farmers and other agricultural producers (Revoredo-Giha and Renwick 2016, Hoyweghen et al. 2021), developing-world/ megacity retail (Boulaksil and Belkora 2017, Ge et al. 2021, Iyer and Palsule-Desai 2019a, Fransoo et al. 2017), and supply chains involving informal laborers in industries such as clothing (Sinha et al. 2020) and recycling/waste management (Tuori 2012).

Informal supply chains are often framed as those involving members of the “informal economy”. Broadly, activities in this category are economically meaningful but do not add to GDP and tax revenue because they are not formally recorded through registration of the involved workers or firms (Delechat and Medina 2020). According to several major development organizations, dominant features of membership in the informal economy include failure to pay taxes, limited productivity, limited access to formal financial institutions, lack of formal employee contracts, absence of benefits and other employee protections, and a focus on survival rather than maximizing profit (Delechat and Medina 2020, OECD and ILO 2019, WB 2021, ILO 2020). It is widely noted that precise definitions differ from country to country due to variations in tax and labor regulations.

However, not all supply chains that feature hallmarks of informality are located in emerging markets, or involve low-income workers. ILO (2020) mentions that members of

the informal economy include both “all workers of the informal sector and informal workers outside the informal sector”. While workers in the former category can be associated with low productivity, limited technology, and reduced contribution to the broader economy, emerging methods of organizing work have led to the increasing prevalence of informal employer-employee relationships outside of the informal sector, and often within (or as contributors to) firms that are considered “formal” from a development perspective. For example, flexible and distributed relationship structures such as spot markets, gig-based markets, outsourcing, and distributed-task platforms meet these criteria.

More generally, informal supply chains can be thought of as supply chains in which traditional relational and structural constraints have been relaxed. We adopt this definition for the purposes of this dissertation. Table 1 provides examples of the ways in which the structure of traditional supply chain functions might be relaxed in informal supply chains.

Supply chain function	Examples of relational/structural constraints present in formal SCs	Potential sources of relaxation
Manufacturing	Quality-controlled batch production	Manual/variable production
Procurement	Contracts	Spot market, crowdsourcing
Order processing	Tech-enabled order mgmt. systems	Informal record-keeping
Warehousing	Cost-minimizing inventory decisions	Capacity-based inventory decisions
Distribution	Optimal lot sizes, fixed schedules	Small, inefficient order volumes
Payment	Accts. payable/receivable, formal credit	Cash, informal credit
Maintenance	Warranties	Informal support without warranties

Table 1: Informal supply chains as a relaxation of traditional supply chains

The relaxation of these constraints can have important implications for supply chain structure and performance; for example, the absence of formal contracts often means that a wider range of actors are able to contribute to informal supply chains. It is thus of both academic and practical relevance to explore new ways of making informal supply chains more productive and beneficial to involved parties. This dissertation contributes to the study of informal supply chains by studying three papers in two distinct industry contexts: 1) informal retail supply chains in emerging markets, and 2) widely distributed digital supply chains enabled by crowdsourcing contests.

1.1 Informal retail supply chains

Informal retailers are often referred to in the literature using alternative names such as “nanostore” or “microretailer”. Fransoo et al. (2017) describe nanostores as “small, typically less than 100 m^2 ...family-operated stores that function as a single establishment” (as opposed to a chain), and which offer relationship-based credit, sell a smaller and less diverse range of items, serve fewer customers, rely on limited technology (e.g., personal devices), sell usage-based package sizes, and (sometimes) sell at higher prices due to inefficient distribution. Informal retailers are thus distinguished from formal retailers, which—in addition to official registration and tax payments—may be characterized by “branded store formats...central ownership or [franchising], with centralized purchasing and distribution functions”, formalized credit, professional employees, larger volumes/diversity of products, larger consumer bases, and enterprise-level technology. Informal retailers can coexist with traditional retail channels, and in fact researchers expect nanostores to retain their dominant status in certain emerging markets given the geographic and socioeconomic realities in developing countries (Fransoo et al. 2017).

Given that traditional supply chain processes are often replaced by relationship-based processes for informal retailers, modifications to relationships with informal retailers could plausibly affect retailer performance. We test this hypothesis empirically in Chapter 2 by evaluating the impact of a disruption to the relationship between a distributor and retailers.

1.2 Crowdsourcing: widely distributed digital supply chains

In the economics literature, contests are broadly defined as settings in which an agent exerts costly effort in order to try and obtain a reward. This effort may consist of actual monetary payment, but commonly takes the form of labor (Dechenaux et al. 2014).

Crowdsourcing is an increasingly common way for a large, otherwise unconnected group of people contribute to a common goal. Crowdsourcing is often facilitated by web-based platforms structured around freelancing, contests, direct payment for microtasks, or competitive programming (ILO 2021). Crowdsourcing contests share several characteristics with

the classic contest structures mentioned in the contest theory literature (see Dechenaux et al. (2014) and Segev (2020) for a review), which are distinguished by their methods of allocating prizes.

Two key elements characterize most contests. First, there is typically a pre-established method of evaluating the quality of participants' effort. Second, this quality evaluation is translated into a determination of who receives prizes via a "contest success function", or prize allocation mechanism. (In some contests there is no need for a quality evaluation metric, as the contest success function maps directly from effort to winner selection.)

Chapter 3 of this dissertation will evaluate two alternative contest evaluation metrics, and Chapter 4 will compare two different prize allocation mechanisms. These chapters aim to shed light onto how workers on distributed-task platforms respond to incentives.

1.3 Dissertation overview

The rest of this document is structured as follows.

Chapter 2 focuses on a distribution network set in the Indian retail industry. The distributor serves as an intermediary between a relatively small number of manufacturers and a much larger number of retailers, some of which are "informal" using a definition based on tax registration. We use causal inference methods to determine the effect of a treatment consisting of disruption to the relationship between the distributor and formal/informal retailers. The key contributions of this chapter are 1) the quantification of the role of trust/relationships in informal vs. formal settings and 2) a focus on durable goods, which are understudied in the literature compared to *e.g.*, fast-moving consumer goods (FMCGs).

In Chapters 3 and 4, the setting is a technology-driven crowdsourcing platform with a massive number of suppliers serving a relatively small number of customers. The platform serves as an intermediary, and its relationships with suppliers are informal due to *e.g.*, a lack of contracts connecting parties. Using data from contests run on this platform, each of these two chapters examines one of the two processes that characterize contests (quality evaluation and prize allocation). In Chapter 3, we hold the prize allocation mechanism constant and study the effect of changing the evaluation metric used to rank participants. We contribute

to the literature on contests by studying performance streaks (which are more commonly used as a gamification tool) as an evaluation metric. In Chapter 4, we hold the evaluation metric constant and change the prize allocation mechanism in a way that shifts the source of earnings uncertainty from the probability of winning a prize to the amount at stake. We contribute to the literature on decision-making under risk by studying the impact of a less-common prize allocation mechanism (evenly split prizes subject to a performance threshold) relative to standard rank-score contests.

Chapter 5 concludes by reflecting on the significance of our insights for the future of informal supply chains.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

The value of long-term relationships when selling to informal retailers - Evidence from India

Olumurejiwa A. Fatunde

Center for Transportation and Logistics, Massachusetts Institute of Technology

Andre P. Calmon

Scheller College of Business, Georgia Institute of Technology

INSEAD

Joann F. de Zegher

Sloan School of Management, Massachusetts Institute of Technology

Gonzalo I. Romero

Rotman School of Management, University of Toronto

2.1 Abstract

Attempts to distribute durable, life-improving goods to customers at the Base of the Pyramid (BoP) – the more than three billion customers who live on less than US\$2.50/day – through traditional supply chains or e-commerce have struggled to succeed at scale. One

hypothesis for why distributors struggle to scale last-mile distribution is poor relationship management with small informal retailers, who are the primary source of retail purchases for BoP customers. These retailers are often embedded within communities, where local and long-term relationships are particularly important to business transactions. We provide empirical evidence for this hypothesis through an analysis of panel data from a distributor selling to 331 formal retailers and 493 informal retailers in India from April 2016-December 2019. Specifically, we study the role of long-term relationships in selling durable goods to informal retailers, by leveraging a staged natural experiment that allows us to examine the effect of a sales agent reallocation on subsequent orders placed by informal and formal retailers. Using two different quasi-experimental methods, we find that formal retailers experience an average performance decrease of at least 35.7% relative to predicted order value and then recover within three sales cycles of a sales agent reallocation; in contrast, informal retailers experience an average performance decrease of at least 70.4% relative to predicted order value, and do not experience sustained recovery within five sales cycles of a sales agent reallocation. This indicates that business relationships, and disruptions to these relationships, are particularly important when selling to retailers in informal markets.

2.2 Introduction

The more than three billion people who live on less than US\$2.50/day, a major part of the Base of the Economic Pyramid (BoP), exhibit distinct purchasing behavior due to the nature of their daily economics. They also have many unmet basic needs; meeting the BoP demand for durable goods such as cookstoves, solar lanterns, and farming equipment is considered key to achieving nearly all of the UN Sustainable Development Goals. These needs have spurred significant research, investment, and human effort focused on developing life-improving technologies tailored to the needs of BoP consumers (see *e.g.*, D-Lab 2021, Hand et al. 2020).

While these efforts have generated a wide range of innovative durable goods, building distribution channels that are effective at getting these durable products to BoP consumers remains a challenge (see, *e.g.*, Baquié and Urpelainen 2017, Deloitte 2017, Gomes and Shah 2018, Aklin and Urpelainen 2020).

BoP consumers often purchase products through small informal retailers, which are part of the informal (or grey) economy – the portion of the economy that is neither taxed nor formally regulated and therefore lacks formal institutions (Webb et al. 2010, Jue 2015). Consistent with the recent literature, our fieldwork indicates that two key characteristics make informal retailers particularly effective venues for durable-good sales in the BoP. First, informal retailers are flexible enough to accommodate BoP customer needs, offering services such as flexible home delivery, the ability to purchase products on credit, and the ability to return products whether or not a warranty agreement is in place (see also, *e.g.*, Viswanathan et al. 2010). Second, informal retailers serve as a key and rare source of information for BoP customers about the value and benefits of new products (see, *e.g.*, Viswanathan et al. 2010, Nozaki 2018, Iyer and Palsule-Desai 2019b). Both of these characteristics are unique to informal retailers because they are deeply embedded within their communities, they are physically close to their customers, and the owners often know their customers by name.

To increase the reach of life-improving technologies tailored to the needs of BoP consumers, it is therefore critical to work with small informal retailers. However, profitably distributing products to these retailers is rife with logistical, financial and behavioral challenges, particularly those related to trust (Garrette and Karnani 2010, Shukla and Bairiganjan 2011, Simanis 2012, Jue 2015). While there is a growing literature on the logistical and financial challenges at the BoP (*e.g.* Boulaksil and Belkora 2017, Boulaksil et al. 2019, Calmon et al. 2017, Gui et al. 2019, Acimovic et al. 2020, Anderson et al. 2018), this paper focuses on understanding the role of trust and relationships in BoP supply chains.

The key liaison between distributor and informal retailer is the sales executive (often referred to as a “sales agent” elsewhere in the literature). Sales executives have varying responsibilities, including some or all of the following: offering and selling products to retailers, negotiating prices, replenishing inventory, providing marketing material, and doing product demonstrations. During our fieldwork with a BoP distributor, it became clear that while there is significant heterogeneity in retailer practices, all retailers have a particular element in common: the relationship between sales executive and retailer is a key repository of information and trust. For example, this relationship critically determines a retailer’s perception of the distributor’s reliability and responsiveness to return or repair requests.

We posit that the relationship of sales executive with retailers is disproportionately important when distributing to informal retailers, which have reduced access to formal institutions and infrastructure. For example, anecdotal reports from our partner distributor’s ground staff suggest that informal retailers suffer a greater disruption in sales than formal retailers following a reallocation of sales executives. Previous literature also provides evidence related to this hypothesis. For example, in a qualitative study of small retailers embedded in the communities of Chennai, India, Viswanathan et al. (2010) indicate that business owners “rely exclusively on interpersonal relationships and the commitments that develop within them to sustain their business," and assert that the influence of relationships is uniquely large in the BoP setting. Using survey-based data, Graça et al. (2016) find that commitments to suppliers are more heavily influenced by relationship-driven factors for buyers in emerging markets like Brazil than for buyers in the US, who are more attentive to the functional benefits of supplier contracts.

To study this hypothesis, we leverage a rich panel dataset (spanning April 2016 - December 2019) from our partner distributor, *Essmart*, which distributed 215 unique durable goods across six product categories to 824 retailers in India at the time of data collection. The distributor sells to both formal retailers and informal retailers (which includes a class of retailers locally known as “kiranas" in India); the relationships with each retailer are managed by 63 full-time sales executives. At any point in time, a single sales executive is assigned to a given retailer.

We use quasi-experimental methods to study the impact that a change in sales executive has on the orders placed by retailers to the distributor, and we examine how this impact differs for formal versus informal retailers. We use two complementary quasi-experimental methods, one based on within-unit matching (a before-after analysis) and one based on between-unit matching/weighting (a difference-in-difference analysis), in order to confirm the robustness of our results.

We find that both formal and informal retailers experience a decrease in order value and diversity immediately following a change in sales executive allocation; with formal and informal retailers experiencing an average decrease of at least 35.7% and 70.4%, respectively, relative to expected order value during the period immediately following a transition. Our

most important finding — supported by both the within-unit and between-unit methods — is that formal retailers recover from this decrease within three sales cycles, whereas informal retailers do not seem to recover sustainably within five sales cycles following the transition. Given the limitations on sample size as the number of sales cycles increases, it is difficult to estimate with confidence how long it takes for informal retailers to recover.

These results provide evidence that the sales executive relationship is disproportionately important for distribution to informal retailers. A disruption to this relationship, in the form of changing the assigned sales executive, has a significantly negative and long-term impact on orders placed by informal retailers. This suggests an important link between operations management at the BoP and the notion of “embeddedness” in organizational behavior (Granovetter 1985); social relationships, such as the relationship between a sales executive and informal retailers, support economic activity in the informal economy, where formal institutions are absent.

The remainder of the paper is structured as follows: Section 4.3 describes the setting and available data, Section 4.4 describes the quasi-experimental methods used in this paper, Section 4.5 presents and discusses the results obtained using each method, and Section 4.6 provides further discussion and concludes the paper.

2.3 Data and Key Definitions

The context for this study is a supply chain with three tiers: (i) a distributor (*Essmart*), (ii) retailers that sell the distributor’s products (among other products), and (iii) customers who purchase products from the retailers. The retailers are located in India’s Tamil Nadu and Karnataka states (see Figure 1).

The retailers operate small- to medium-sized shops with limited storage space, and thus carry limited inventory of the distributor’s items; for bulky items, like cookstoves, retailers often carry only one unit. As a result, the distributor has established a network of distribution centers from which retailers’ inventory can be replenished with short lead times (typically one or two days). Retailers therefore typically use the distributor for “deliver-to-order” sales; they carry a demonstration unit inside the store and place an order with the distributor only

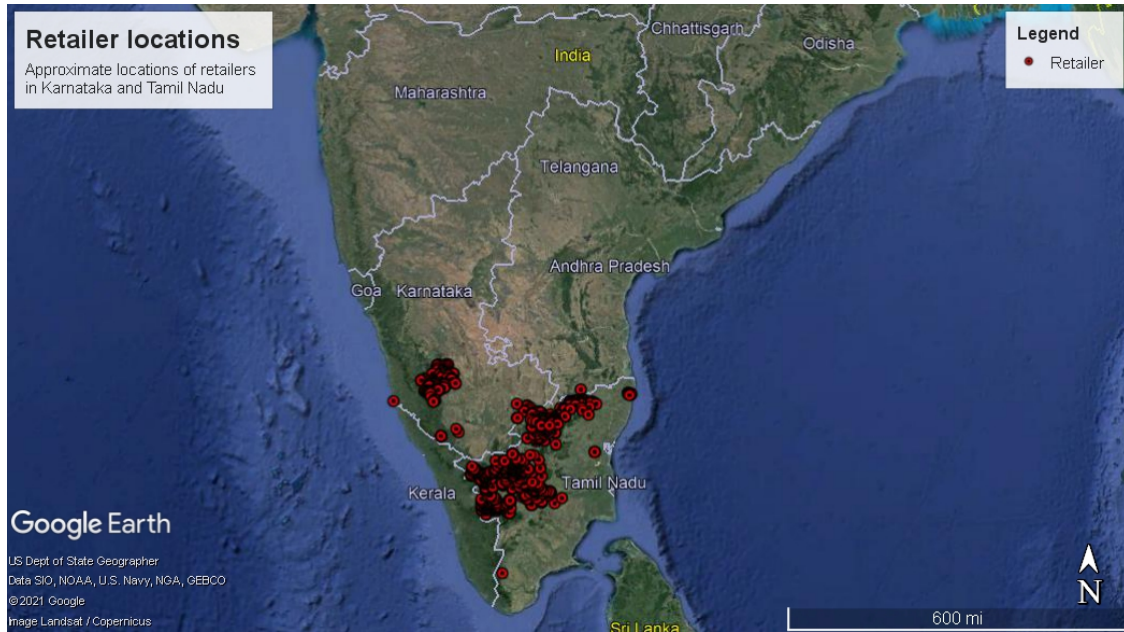


Figure 1: Retailer locations

when a customer wishes to purchase the item. When a retailer places a purchase order with the distributor, the items in the order are delivered by a sales executive.

We obtained three datasets from the distributor: orders placed by retailers, retailer tax registration information, and sales executive employment information. In addition, we collected data on time-invariant retailer characteristics. An anonymized version of the data and our code is available at [link redacted for peer review].

We converted this data into a panel dataset, where the unit of observation is a retailer-cycle. The length of a retailer’s “cycle” is defined as the median interval, in days, between consecutive orders from that retailer (our results are robust to using retailer-month as the unit of observation for applicable methods). We use cycles as a standardized unit of time in order to compare performance across a wide range of ordering policies that retailers may adopt (*e.g.* due to proximity to a distribution center, inventory space, etc.). To give a concrete example; it would be difficult to compare the impact of an intervention on the performance of two retailers in a given month when one retailer uses a periodic review model with a review period of one month and the other retailer uses a review period of six months; instead, we would want to compare orders over their respective inter-order intervals. Basing our analysis on observed inter-order intervals achieves this.

Retailers start ordering from the distributor at different points in time, and each retailer enters the dataset on the date on which it placed its first order. The panel is therefore unbalanced. The resulting panel has 14,102 retailer-cycles across 616 retailers (208 out of the 824 retailers placed only one order with the distributor and therefore have an undefined inter-order interval; see also 2.3.1).

We next define each of the datasets and key definitions used in this study.

2.3.1 Retailer order data

We define an order as a positive number of SKUs purchased by a retailer on a given day. The order data includes 5,538 orders placed by 824 retailers on 967 days during the overall period of 1,358 days (approximately 195 weeks) between April 2016 - December 2019. The order data includes the date of the order, the order value, the categories to which the products in the order belong, the ID of the retailer that placed the order with the distributor, and the ID of the sales executive who handled the order. The product categories sold by the distributor are listed in Table 1.

Product category	Total volume
Kitchen	44,106
Farming	11,968
Lighting	6,536
Health & Safety	1,670
Home Comforts	646
Gadgets	145

Table 1: Product categories sold by the distributor (April 2016-December 2019)

We define order diversity as the number of product categories included in a given retailer order. If more than 90% of a retailer’s order volume (over the entire time horizon) consists of products within a single product category, we say that the retailer is a “specialist” in that category.

Out of the 824 retailers (with a total of 5,538 orders) uniquely identified in the data, we excluded 208 retailers that placed only one order over this time period because their inter-order interval is undefined. These exclusions left us with 616 retailers that placed 5,330 orders.

Because a single sales executive is associated with each order, the order data allows us to infer whether a new sales executive has been assigned to a given retailer. Note that while we know *whether* a retailer has witnessed a transition in sales executive between two orders, we cannot infer *when* exactly the new sales executive was assigned. Retailers in our sample experienced between zero and seven sales executive transitions (Figure 2). More than half of the retailers in the sample never experienced a transition; 125 experienced a single transition.

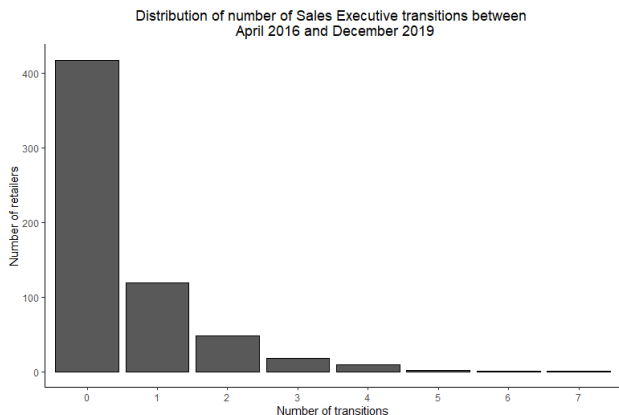


Figure 2: Distribution of number of transitions across retailers.

In the analyses that follow, we focus on retailers that have either no transition or exactly one transition. This ensures that the estimated effect of a transition is not contaminated by subsequent transitions. Of the 616 remaining retailers, we therefore further excluded 73 retailers with two or more transitions, leaving us with a base sample of 543 retailers that witnessed at most a single transition. Of these retailers, we excluded 65 that experienced a transition in their first cycle, since we require at least one pre-transition period and one post-transition period in order to complete a before-after comparison. The resulting dataset contains 60 treated retailers that experience exactly one sales executive transition and 418 candidate control retailers that never experience a transition.

A summary of the exclusion criteria described above is provided in Table 2. Table 2 also includes a breakdown of retailers by formal/informal status, which we define next using data on retailer tax registration.

	<i>Subgroup</i>		
	All	Formal	Informal
# unique retailers served by distributor during study period	824	331	493
# retailers with >1 order	616	263	353
# retailers with 0 transitions	418	159	259
# retailers with exactly 1 transition	125	67	58
# retailers with 1 transition after first period (" <i>treated retailers</i> ")	60	36	24
# retailers with 0 transitions (" <i>candidate control retailers</i> ")	418	159	259

Table 2: Breakdown of retailers in our sample

2.3.2 Retailer tax registration

Businesses in India are required to register for a Goods and Services Tax (GST) number if their aggregate annual turnover surpasses a certain threshold (thresholds differ for goods and services), if they operate across multiple states or internationally, if they participate in e-commerce, or if they were registered under older tax regimens such as VAT (ClearTax 2021b). For most states, including the distributor’s states of operation, the initial turnover threshold (beginning in July 2017) was INR 2 million (\approx \$26,508), and it increased to INR 4 million (\approx \$57,720) in April 2019 for businesses whose only activity is to supply goods, such as most of the retailers in the distributor’s network (ClearTax 2021a).

Businesses whose turnover falls below the threshold may voluntarily register for a GST number in order to capitalize on several benefits. For example, voluntary registration may increase retailers’ access to a wider range of suppliers; some suppliers will not sell large volumes of goods to unregistered customers, since doing so may attract unwanted scrutiny from regulators. GST registration may also provide greater access to large amounts of capital, because some banks require GST registration as a requirement for business loans.

Given the structure of the regulations governing GST registration, we can have confidence that retailers with *no* GST number (1) have an upper bound on annual revenue, (2) operate within a single state (and in many cases, a much smaller area), (3) do not sell goods online, and (4) do not charge their customers GST. Our partner distributor also indicated that many of the unregistered retailers are small family businesses run by a sole operator, or perhaps with a small number of helpers. These characteristics align with those typically ascribed to informal retailers in the literature.

	Formal	Informal	<i>p</i> -value
Mean # of full-time staff	2.74	1.36	0.025
Mean approximate floor area in sq. ft.	785	461	0.011
Mean max. retail price, most expensive product (INR)	17,570	11,923	<0.001
Technology: % retailers relying on memory to manage sales & inventory	11	23	0.02

Table Notes: *p*-values are computed using the non-parametric Kruskal-Wallis test (for continuous variables) or a chi-squared difference in proportion test (for variables representing % of retailers). We used non-parametric tests due to skewness in the data.

Table 3: Comparison of survey results for 127 “informal” and 125 “formal” retailers.

We therefore classify retailers with no GST number as “informal retailers” and retailers that have a GST number as “formal retailers.” Using survey data (described in Section 2.3.4), we further micro-found this classification as a valid approximation of other metrics of retailer “formality”; retailers classified as “informal” have significantly fewer full-time staff, significantly smaller floor area, significantly less expensive products, and significantly higher reliance on memory to manage sales and inventory (see Table 3).

Note that this definition implies that we might assign some retailers with informal characteristics to the “formal” category, since some informal retailers might voluntarily obtain a GST number—which would cause them to fall into the formal category based on our criteria. As a result, we might *underestimate* the difference between treatment effects for formal versus informal retailers. In contrast, a retailer classified as “informal” in our data is unlikely to actually be a formal retailer, given the penalties associated with failing to register for a GST number for companies that are required to do so.

2.3.3 Sales executive employment

We also obtained data on the employment history of sales executives. For each sales executive, the data contains the unique sales executive ID, the date of employment, date of departure, and current active status.

2.3.4 Retailer-level characteristics

Finally, we collected supplementary information on retailer-level characteristics. First, we obtained information from the distributor on the GPS coordinates or address of each re-

tailer, which we used to compute the driving distance (in km) between a retailer and the corresponding distributor’s branch.

Second, we conducted a survey in April 2019 on 251 of the retailers (127 retailers classified as informal and 124 retailers classified as formal). The survey was conducted by sales executives during their routine interactions with the retailers over a period of approximately three weeks. As such, only retailers that were actively placing orders with the distributor at the time of the survey were included. Information collected through the survey includes number and type of employees, technology use, type and price range of products sold, and retailer size and layout.

2.4 Methodology

We use quasi-experimental methods to test the impact of a sales executive transition on the performance of formal versus informal retailers. The treatment of interest is therefore the assignment of a new sales executive to a retailer.

We use the dummy indicator $D_{i,t}$ to denote whether or not retailer i received treatment in sales $t \in [1, T]$, where T is retailer-dependent given the unbalanced nature of the panel. Retailer i is initially assigned a treatment indicator of zero ($D_{i,t} = 0$) starting from the retailer’s first ($t = 1$); the indicator is incremented to one during the treatment t_0 in which a new sales executive is assigned ($D_{i,t_0} = 1$), and it remains at one thereafter for all $t \geq t_0$. Because we limit the scope of our analysis to retailers that experience a single transition, the treatment condition is both non-reversible and non-repeatable.

We aim to estimate the causal effect for each subgroup during a specified post-treatment period. Let i denote a retailer and \mathcal{G} denote the subgroup of retailers under consideration, *i.e.* $\mathcal{G} \subset \{\mathcal{I}, \mathcal{F}\}$, where \mathcal{I} is the set of informal retailers and \mathcal{F} is the set of formal retailers. We let $Y_{i,t}(1)$ denote the outcome of retailer i in period t where $D_{i,t} = 1$, and let $\hat{Y}_{i,t}(0)$ denote the estimated counterfactual outcome of retailer i in period t . The group-time average treatment effect for subgroup \mathcal{G} in cycle $t_0 + F$ is given by:

$$\hat{\tau}_{F,\mathcal{G}} = \mathbb{E}[Y_{i,t_0+F}(1) - \hat{Y}_{i,t_0+F}(0)|i \in \mathcal{G}]. \quad (2.1)$$

Because our eventual aim is to compare *relative* effects of two subgroups with different baseline performance levels, the relevant quantity of interest is the ratio of potential outcomes, rather than the difference (see *e.g.*, Morgan and Winship (2015) for others who followed this approach):

$$\hat{\tau}'_{F,\mathcal{G}} = \mathbb{E} \left[\frac{Y_{i,t_0+F}(1)}{\hat{Y}_{i,t_0+F}(0)} \middle| i \in \mathcal{G} \right]. \quad (2.2)$$

The key decision is how to define the counterfactual “control” outcome for each treated unit at the time of interest, *i.e.* $\hat{Y}_{i,t}(0)$. There are two options for members of the comparison group: (1) observations from the same retailer during the pre-treatment period (*within-unit matching*), and (2) observations from other retailers that are “similar enough” on defined characteristics (*between-unit matching/weighting*). We next define estimators based on each of these potential comparison groups.

2.4.1 Within-unit matching

Within-unit matching involves using the pre-treatment outcomes for retailer i to estimate $\hat{Y}_{i,t}(0)$. This method therefore effectively compares the outcomes of each retailer before and after treatment, and we use a nonparametric before-after matching estimator of Average Treatment Effect on the Treated (ATT) to estimate the causal effect. Figure 3 shows, for each subgroup, the time at which each of the treated retailers received treatment. By comparing treated and control units across time within the same unit, this estimator has the benefit of controlling for potential unobserved unit-specific, time-invariant confounders by capturing unit-level heterogeneity.

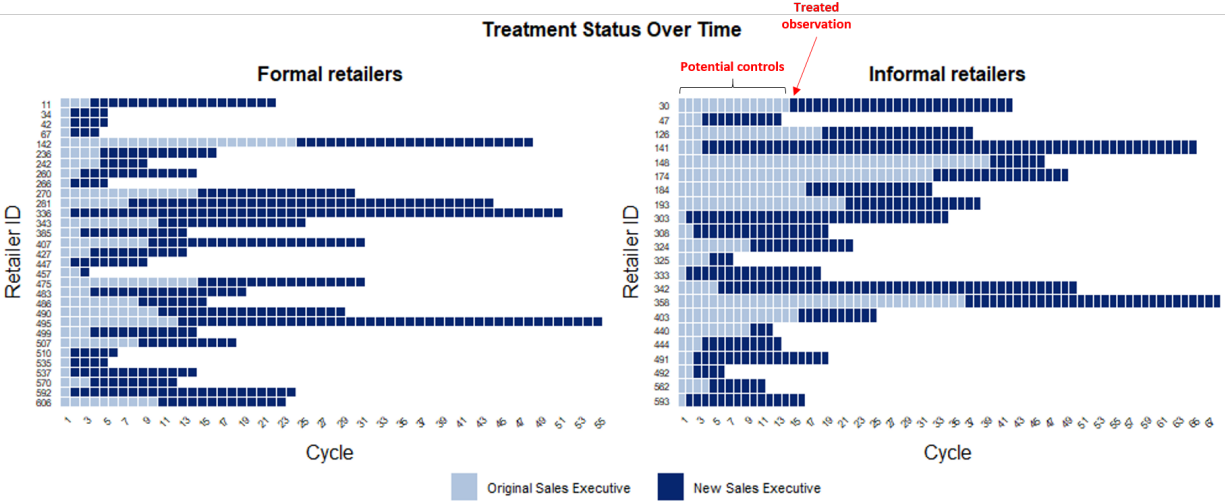


Figure 3: Time at which treated retailers received treatment, and potential controls using within-unit matching.

Imai and Kim (2019) develop a non-parametric within-unit matching estimator that compares an observed post-treatment outcome with the mean of the pre-treatment outcomes for a given unit. They demonstrate that this estimator is equivalent to the conventional inverse propensity-score weighted linear regression with unit-fixed effects, but allows relaxing the linearity assumption of linear regression, and does not require assumptions about the distribution of outcomes. This relaxation of these traditional assumptions through a non-parametric approach is important in our context, given the limitations on our sample size.

The within-unit method requires three key identification assumptions. First, that there are no unobserved time-varying confounders. Second, that past treatment has no effect on current outcome. Third, that past outcomes have no effect on current treatment. Figure 3 indicates significant diversity in the timing of treatment across retailers (this is also true when using calendar months rather than cycles as the time unit). Any time-varying confounders would therefore add noise that would reduce our ability to detect a significant treatment effect. The second assumption, also known as the absence of a “carryover” effect, is met for each lag-specific estimate because treatment occurs just once. The third assumption would be violated if, for example, there were a pre-determined threshold of performance that triggered a transition (*e.g.*, a dollar amount or minimum number of products). While this is not the case, general sales executive performance is a factor in re-allocation decisions.

We therefore also use *between-unit matching/weighting* (described in Section 2.4.2) and find that time-varying confounders are unlikely to significantly bias our analyses; the two methods yield consistent results.

Using Imai and Kim (2019), for each subgroup \mathcal{G} and cycle $t_0 + F$, the (within-unit) estimated contemporaneous treatment effect is:

$$\hat{\tau}_{F,\mathcal{G}} = \mathbb{E}[Y_{i,t_0+F}(1) - \bar{Y}_i(0) | i \in \mathcal{G}],$$

where $\bar{Y}_i(0)$, the “control mean” of retailer i , is given by:

$$\bar{Y}_i(0) = \frac{\sum_{t=1}^{t=t_0-1} Y_{i,t}}{t_0 - 1}.$$

The relative form of the estimator is:

$$\hat{\tau}'_{F,\mathcal{G}} = \mathbb{E} \left[\frac{Y_{i,t_0+F}(1)}{\bar{Y}_i(0)} \middle| i \in \mathcal{G} \right].$$

Note that the magnitude of the effect in post-treatment cycle $t_0 + F$ as a percentage decrease from mean pre-treatment performance is given by $1 - \hat{\tau}'_{F,\mathcal{G}}$.

To calculate confidence intervals, we use a block bootstrap procedure, which is often used when the assumptions needed to accurately estimate closed-form uncertainty measures cannot be relied upon (*e.g.*, Gareth James 2013, Hazlett and Xu 2018, Liu et al. 2020). In this case, the ratio format of our outcome makes it difficult to make distributional assumptions. We therefore sample with replacement 1,000 times at the retailer level for retailers in subgroup \mathcal{G} , compute each lag-specific treatment effect for each of the 1,000 samples, and then use percentile methods to identify confidence intervals for the treatment effect corresponding to each post-treatment cycle.

Our identification strategy allows us to recover a consistent and unbiased estimate of the ATT in each post-treatment cycle, computed separately for formal and informal retailers. However, we note that the number of retailers included in the within-unit analysis shrinks over time, since we do not have observations for all retailers across all post-treatment cycles (see also Figure 3). When reporting our results, we therefore display the number of

observations included.

In a pre-processing step, we remove outliers in the data. We use a threshold rule to define how extreme an observation must be to be considered an outlier; for a given variable, we remove retailers with observations that fall more than $X \cdot IQR$ (Inter-Quartile Range) above the 75th percentile of that variable in at least one period between $t_0 - 5$ and $t_0 + 5$. In Section 2.4.2, outliers based on relative order value and relative order volume (as defined by Y_{i,t_0+F} in Section 2.4.2) have an important impact on achieving balance between control and treatment retailers; both are thus considered when removing outliers. For consistency between the two methods, we remove retailers using the same criterion for the within-unit method. For formal retailers, we set $X = 17$ and $X = 15$ for order value and order volume, respectively. For informal retailers, we set $X = 10$ and $X = 6$ for order value and order volume, respectively.

2.4.2 Between-unit matching/weighting

The within-unit estimator in Section 2.4.1 controls for unobserved retailer-specific, time-invariant confounders, but assumes that there are no time-varying confounders and that past outcomes have no effect on treatment. Here, we propose an estimator that matches treated retailers to control retailers in order to control for potential time-varying and time-specific confounders. We make the following five identification assumptions. First, we assume no spillovers (*i.e.*, retailers are not affected by the treatment status of other retailers). Second, we assume no anticipation effects in pre-treatment periods. Third, we assume conditional parallel trends after controlling for treatment, outcome, and covariate history in the specified lag periods (which here refers to the single period before treatment). Fourth, we assume sufficient common support (*i.e.*, there is a nonzero possibility of treatment in each period included in the horizon). Fifth, we assume no unobserved time-invariant confounders after controlling for covariate history and treatment. These assumptions, together with the irreversibility of treatment, are sufficient to identify the causal effect of interest even in the absence of random sampling.

Imai et al. (2020) extend the non-parametric within-unit matching technique described in Imai and Kim (2019) to allow for comparison of treated units with (never-before-treated)

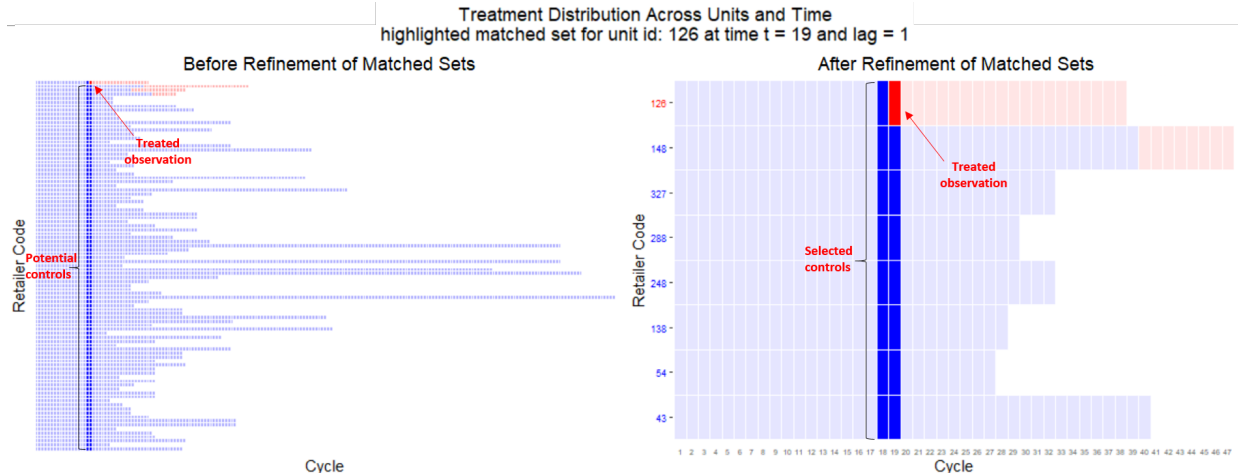


Figure 4: Treated retailer and control retailers for between-unit k -nearest-neighbor matching estimator with all potential control units (left panel) and with matched control units only (right panel).

control units. They do so by comparing a specific retailer-period observation from a treated retailer to one or more control retailers that match the treated retailer (with a degree of similarity that must be specified) on a set of covariates. This includes both observations from control retailers that are never treated throughout the entire time horizon and pre-treatment observations from treated retailers whose treatment is introduced *after* the retailer under study. Figure 4 provides an example of a retailer with all potential control units, as well as with matched control units only.

We use two different approaches to identifying the set of control retailers included in ATT estimation under this method. The first approach matches treated retailers to the k candidate control retailers that are most similar to the treated retailer during the cycle preceding treatment (*k-nearest neighbor matching*). The second retains all candidate control retailers and weights them based on similarity to treated retailers (*inverse propensity score weighting*). In order to study the effect of treatment from period t_0 to $t_0 + F$, we require that retailers selected as controls for retailer i do not experience treatment in any period $t \in [1, t_0 + F]$.

We use the non-parametric estimator introduced in Imai et al. (2020) for the ATT of stable policy change (*i.e.*, irreversible treatment) relative to no policy change. This is a difference-in-difference (DiD) estimator for a particular time period. The estimator for the

treatment effect F after a transition in t_0 for subgroup \mathcal{G} is given by:

$$\hat{\delta}_{F,\mathcal{G}} = \frac{1}{\sum_{i \in \mathcal{G}} \sum_{t=2}^{T-F} X_{i,t}} \sum_{i \in \mathcal{G}} \sum_{t=2}^{T-F} X_{i,t} \left\{ (Y_{i,t_0+F} - Y_{i,t_0-1}) - \sum_{i' \in M_i} w_{it}^{i'} (Y_{i',t_0+F} - Y_{i',t_0-1}) \right\}, \quad (2.3)$$

where $X_{i,t} = D_{i,t}(1 - D_{i,t-1}) \times \mathbf{1}\{|M_i| > 0\}$, and M_i denotes all control retailers in the matched set for retailer i . Thus, $X_{i,t}$ is exactly equal to 1 in period t_0 (the treatment period) for retailer i if we can identify one or more suitable controls, and is zero otherwise (Imai et al. 2020). Furthermore, $w_{it}^{i'}$ are non-negative normalized weights which determine the relative influence of each control retailer such that $w_{it}^{i'} \geq 0$ and $\sum_{i' \in M_i} w_{it}^{i'} = 1$. We use the weights developed in Imai et al. (2020) for the ATT of a “stable”, or long-term, treatment; in the case of k -nearest neighbor matching the weights are $1/k$ for each of the k control retailers, and in the case of inverse propensity score weighting, the weights are given by a function of the cycle-specific propensity scores in each of the post-treatment periods being considered:

$$w_{it_0}^{i'} = \prod_{f=0}^F w_{it_0f}^{i'*},$$

where

$$w_{it_0f}^{i'*} \propto \frac{\hat{\pi}_{i,t_0+f}}{1 - \hat{\pi}_{i,t_0+f}}$$

such that $w_{it_0f}^{i'} \geq 0$ and $\sum_{i' \in M_i} w_{it_0f}^{i'} = 1$, and $\hat{\pi}_{i,t_0+f}$ is the estimated propensity score for retailer i in period $t_0 + f$.

The summation indices in the numerator and the denominator of the first term on the right-hand side of (2.3) ensure that we only include treated retailers with at least one period of data before the transition (because period $t_0 - 1$ is used as the reference point for DiD) and at least F periods of data after the transition (because we are computing the treatment effect for period $t_0 + F$).

While Equation (2.3) identifies the absolute treatment effect, we are interested in the relative treatment effect in each subgroup, given the goal of comparing relative effects across groups. The relative variant of the DiD estimator for the ATT is the same as Equation (2.3), but with relative outcomes (defined as Y'_{i,t_0+F} and Y'_{i,t_0-1}) in place of absolute outcomes.

Hence, the equation becomes:

$$\hat{\delta}'_{F,\mathcal{G}} = \frac{1}{\sum_{i \in \mathcal{G}} \sum_{t=2}^{T-F} X_{it}} \sum_{i \in \mathcal{G}} \sum_{t=2}^{T-F} X_{i,t} \left\{ (Y'_{i,t_0+F} - Y'_{i,t_0-1}) - \sum_{i' \in M_i} w_{it}^{i'} (Y'_{i',t_0+F} - Y'_{i',t_0-1}) \right\},$$

where

$$Y'_{i,t_0+F} = \frac{Y_{i,t_0+F}}{\bar{Y}_{i,t_0+F}}, \quad \text{and} \quad \bar{Y}_{i,t_0+F} = \frac{\sum_{t'=1}^{t_0-1} Y_{i,t_0} P_{it'}}{\sum_{t'=1}^{t_0-1} P_{it'}}.$$

Herein, $P_{it'}$ is an indicator variable indicating whether an outcome for retailer i is non-zero in cycle t' . Y'_{i,t_0-1} is defined analogously to Y'_{i,t_0+F} .

For between-unit methods, we can evaluate the magnitude of the treatment effect for each post-treatment cycle $t_0 + F$ by measuring the percentage change between the counterfactual and the observed outcome for that period:

$$\text{Effect magnitude}_{F,\mathcal{G}} = \mathbb{E} \left[\frac{Y'_{i,t_0+F}(1) - \hat{Y}'_{i,t_0+F}(0)}{\hat{Y}'_{i,t_0+F}(0)} \middle| i \in \mathcal{G} \right],$$

where the unit-specific counterfactual is given by:

$$\hat{Y}'_{i,t_0+F}(0) = \sum_{i' \in M_i} w_{it}^{i'} Y'_{i',t_0+F}.$$

Standard errors are computed using the weighted block-bootstrap procedure developed for between-unit matching by Imai et al. (2020). In this procedure, matching is not re-done for each bootstrap replication; instead, sampling is done at the retailer level, and the ATT is re-computed with each retailer-cycle observation in the resulting samples being weighted based on the number of times the observation was used for matching in the original sample. Importantly, these standard errors reflect uncertainty conditional on the matching process that assigns each treated retailer to a set of control retailers. We sampled with replacement 1,000 times at the retailer level and recomputed the treatment effect for the sample produced in each iteration, using the weights as described above.

Next, we describe in more detail how we identify the set of control retailers included

in the estimation, *i.e.* the matched set M_i for retailer i . Imai et al. (2020) include only time-varying covariates to determine which retailers to include in M_i . However, because we match on pre-treatment history for only a single period (cycle $t_0 - 1$), we are able to include both time-varying and time-invariant covariates to determine matches. The covariates that we consider are:

- Retailer age (days between retailer’s first order and first day of the current cycle);
- Average value and volume of orders per period between the retailer’s first order in the dataset and the first day of the current order cycle;
- Sales executive-retailer relationship length (days between assignment of departing sales executive and first day of the current cycle);
- Distance to distributor branch;
- Median and standard deviation of cycle length;
- Dummy variables for branches that have treated/control retailers in both \mathcal{I} and \mathcal{F} .

In addition, to control for seasonality and unobserved time-specific confounders, we consider the following time markers (determined on the first day of each cycle) in matching: season, month, year, and month-year.

For k -nearest neighbor matching, the choice of k (maximum number of controls allowed in the final matched set) is a key parameter. To select k , we iteratively examine the balance of key covariates between treated and control groups by varying the covariates included in the matching, the choice of k , and the outlier threshold X . Matching and balance evaluation may be performed iteratively without biasing the results, as long as outcome values are not consulted along the way (Ho et al. 2007). We stop when we achieve an acceptable level of balance on historical order value and volume, as well as time-varying covariates, which are critical to evaluating the validity of the conditional parallel trends assumption (Imai et al. 2020).

To evaluate balance, we compute the standardized mean difference (SMD), which is a widely used unitless quantity, before and after matching. The SMD for a given covariate j and treated unit i at time period t is given by

$$SMD_{i,t,j} = \frac{x_{i,t,j} - \bar{x}_{i',t,j}}{s_{\text{treated}}},$$

for the original data, and by

$$SMD_{i,t,j}^m = \frac{x_{i,t,j} - \sum_{i' \in M_i} w_{ii'}^j x_{i',t,j}}{s_{\text{treated}}},$$

for the matched data. Herein, $x_{i,t,j}$ and $\bar{x}_{i',t,j}$ refer, respectively, to the value of covariate j for treated unit i and the sample mean of covariate j for the control units matched to unit i . s_{treated} denotes the sample standard deviation for (unmatched) treated units (Ho et al. 2011, Imai et al. 2020).

For each covariate j , the SMD is calculated for each treated unit for the period immediately preceding treatment $t_0 - 1$, and then aggregated across all treated units (for all units and matched units respectively):

$$\bar{B}_{j,t_0-1} = \frac{1}{N_{\text{treated}}} \sum_{i=1}^{N_{\text{treated}}} SMD_{i,t_0-1,j}, \quad \bar{B}_{j,t_0-1}^m = \frac{1}{N_{\text{treated}}} \sum_{i=1}^{N_{\text{treated}}} SMD_{i,t_0-1,j}^m,$$

where N_{treated} is the overall number of treated units (Imai et al. 2020).

For a given combination of outlier thresholds and covariate specification, we evaluate the quality of matching by examining \bar{B}_{j,t_0-1}^m using covariate values from cycle $t_0 - 1$. We refer to \bar{B}_{j,t_0-1} and \bar{B}_{j,t_0-1}^m , respectively, as the pre- and post-matching *aggregated SMD*. An aggregated SMD with an absolute value greater than 0.1 is considered to indicate an unacceptable level of imbalance between groups (Wang et al. 2013).

Figure 5 shows the balance results for key covariates using a range of possible permitted matches.

We choose k separately for each group in \mathcal{G} so as to optimize balance in period $t_0 - 1$, which serves as the reference period for the DiD estimator. We choose $k = 5$ for formal retailers and $k = 7$ for informal retailers, thereby keeping the aggregated SMD below (or close to) the threshold of 0.1 for key covariates of interest (see Figure 5).

Table 4 shows the pre- and post-matching aggregated SMD for covariates that were included in the specifications for either formal or informal retailers, using the values of k

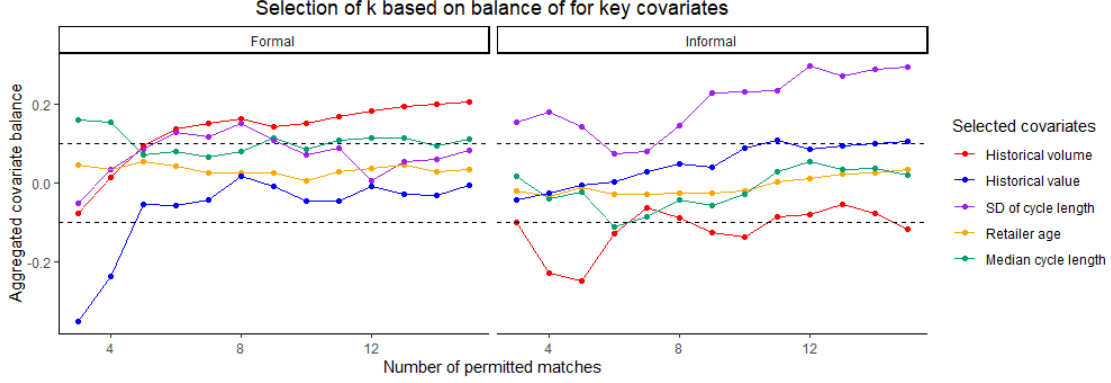


Figure 5: Balance on key covariates by number of permitted matches based on \bar{B}_{j,t_0-1}^m .

mentioned above. We find that treatment and control retailers are balanced after matching on all covariates of interest – as indicated by all \bar{B}_{j,t_0-1}^m being below the threshold value of 0.1. The fact that we achieve balance on historical measures of value and volume allays our concerns about potential selection bias.

	<i>Formal</i>		<i>Informal</i>	
	\bar{B}_{j,t_0-1}	\bar{B}_{j,t_0-1}^m	\bar{B}_{j,t_0-1}	\bar{B}_{j,t_0-1}^m
Historical average value per order	0.06	-0.05	0.29	0.03
Historical average volume per order	0.16	0.09	-0.09	-0.06
Median inter-order interval	0.09	0.07	-0.28	-0.09
Std. deviation of inter-order interval	-0.24	0.09	0.43	0.08
SE-retailer relationship length	-0.00	0.06	0.09	-0.07
Retailer age at cycle start	-0.00	0.06	0.13	-0.03
Distance from retailer to branch	-0.29	-0.06	-	-
Mettupalayam Branch	-0.73	-0.00	-	-

Table 4: Covariate balance before and after k-nearest-neighbor matching using values $k = 5$ for formal retailers and $k = 7$ for informal retailers.

Table 5 provides a summary of the number of retailers included in the treatment and control groups for both inverse propensity score weighting and k -nearest neighbor matching.

Before computing the between-unit ATT estimates, we pre-processed the data by removing outliers. We remove the same set of outliers that were removed for the within-unit matching method, as described in Section 2.4.1.

Between-unit sample sizes

	<i>Formal</i>		<i>Informal</i>	
	Control	Treated	Control	Treated
IPW sample	143 (132/11)	25	229 (219/10)	19
kNN sample	69 (63/6)	25	92 (88/4)	19

Table notes: Number of retailers included to calculate DiD estimates using Inverse Propensity Score Weighting (IPW) and k -nearest neighbor (kNN) matching. The kNN sample is smaller than the IPW sample because only the k most similar control retailers are retained. The control units in these samples include both never-treated retailers (first figure in parentheses) and retailers which are treated in later periods but serve as controls for retailers treated earlier (second figure in parentheses).

Table 5: Number of retailers included after matching for formal and informal retailers

2.5 Results

In this section, we present the results from both the within-unit matching method described in Section 2.4.1 and the between-unit matching method described in Section 2.4.2. We present results for two outcomes of interest: order value and order diversity.

We show the treatment effect for up to five sales cycles after treatment. The number of cycles was chosen to balance three considerations. First, many of the distributor’s performance reviews take place on a quarterly cycle, so a period of five cycles (for the median retailer, equivalent to five to six months) captures at least a full quarter, with some buffer on both sides. Second, after a longer period, retailer performance might be influenced by factors unrelated to the transition. Third, after five sales cycles, the number of observations becomes too small to confidently estimate an effect.

We also show the pre-treatment trend in relative outcomes for up to five cycles preceding treatment in order to demonstrate no effect in pre-treatment periods and, hence, the validity of our control observations.

2.5.1 Within-unit matching

Figure 6 shows the ATT for formal and informal retailers using the within-unit estimator. Note that the null effect for this estimator is one, because the within-unit estimator computes a ratio of observed post-treatment outcomes to potential outcomes under control

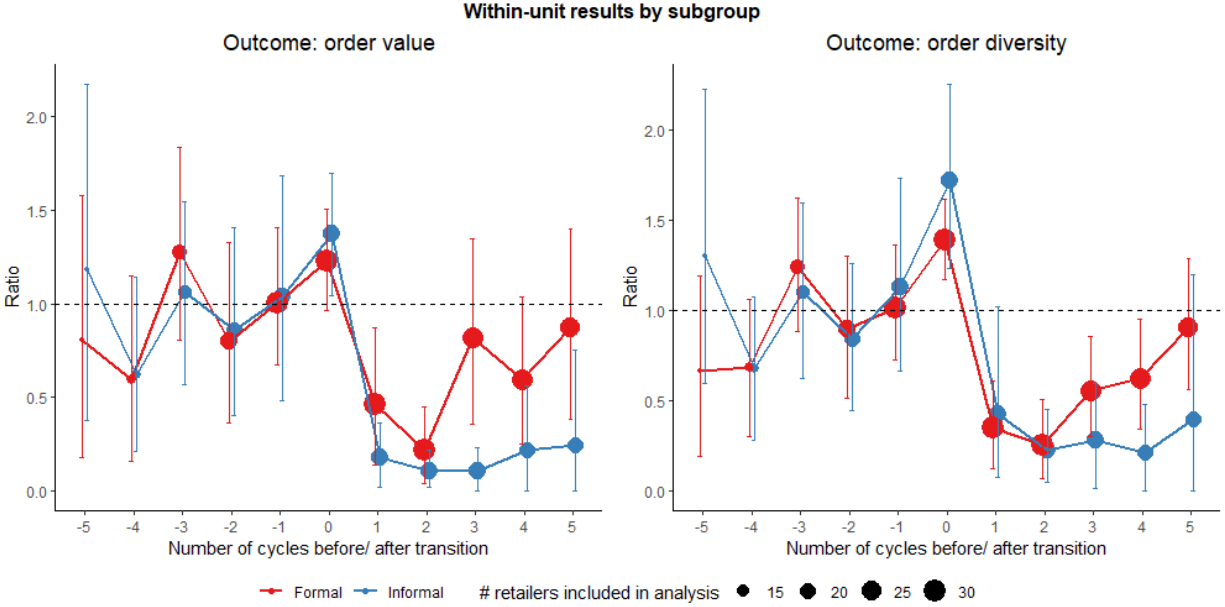


Figure Notes: Pre-treatment performance (cycles -5 to -1), average treatment effect on the treated (cycles 1 to 5), and 95% confidence intervals for formal and informal retailers that experience a change in sales executive. Confidence intervals are computed using block bootstrap with replacement (1,000 replications).

Figure 6: ATT based on the within-unit estimator.

(see Section 2.4.1). A ratio greater than one corresponds to a positive treatment effect, or average post-treatment performance which is better than expected based on pre-treatment performance. Conversely, a ratio between zero and one suggests a negative treatment effect.

Note that retailers experience a significant increase in order value and diversity during cycle t_0 . However, this spike is a methodological artifact: cycle t_0 contains – by construction – an order for all retailers, whereas for other cycles some retailers may have zero orders.

For both groups, the most extreme change to order value and diversity occurs in the sales cycle following a sales executive transition. Comparing the magnitude of the average effect is instructive: for order value, formal retailers experienced a decrease in performance (relative to pre-treatment means) of 53.8%, and informal retailers experienced a corresponding reduction of 81.9%. For order diversity, formal and informal retailers experienced reductions of 64.9% and 57.6%, respectively, in the first post-treatment cycle. These reductions in performance do not differ significantly between both subgroups, but differ significantly from the null effect of one.

The most notable finding is that formal retailers return to their pre-treatment mean

order value within three sales cycles, whereas informal retailers do not recover their mean pre-treatment performance even after five sales cycles. In the appendix, we reproduce these results using months as the time unit rather than cycles. This finding provides evidence for our main conjecture presented in the introduction: the disruption of a sales executive relationship is disproportionately important for informal retailers in the BoP.

In the next sub-section, we verify that this insight is preserved when we control for time-varying confounders using the between-unit matching/weighting method.

2.5.2 Between-unit matching/weighting

Figure 7 shows the ATT for formal and informal retailers using the between-unit estimator. Note that here a null effect is represented by an ATT of zero, because the between-unit estimator is a difference-in-difference estimator. A negative (positive) treatment effect corresponds to worse-than-expected (better-than-expected) performance, where the “expected” performance is defined by the estimated counterfactual outcome under “no treatment”.

Because the difference-in-difference estimator uses cycle $t_0 - 1$ as a reference point, the treatment effect for cycle $t_0 - 1$ is exactly equal to zero, with standard errors of zero.

The treatment effect in all other pre-treatment periods is statistically indistinguishable from zero, suggesting that the sharp decrease in the cycle following treatment is indeed attributable to a sales executive transition.

In the period immediately after a change in sales executive, informal retailers perform an average of 70.4% worse than expected for order value, and an average of 50.0% below expectations for order diversity. In addition, this under-performance is statistically significant. Formal retailers experience more modest reductions of 35.7% and 46.7% for the two outcomes, respectively, and these reductions are also statistically significant.

Consistent with our findings using the within-unit estimator from Section 2.5.1, we observe that formal retailers return to expected performance levels within three sales cycles, whereas there is no consistent evidence that informal retailers return to (and remain at) expected levels. While our hypothesis is that these results are attributable to trust, alternative explanations could explain the observed difference in results between formal and informal retailers. For example, if one or both types of retailers changed the types of products that

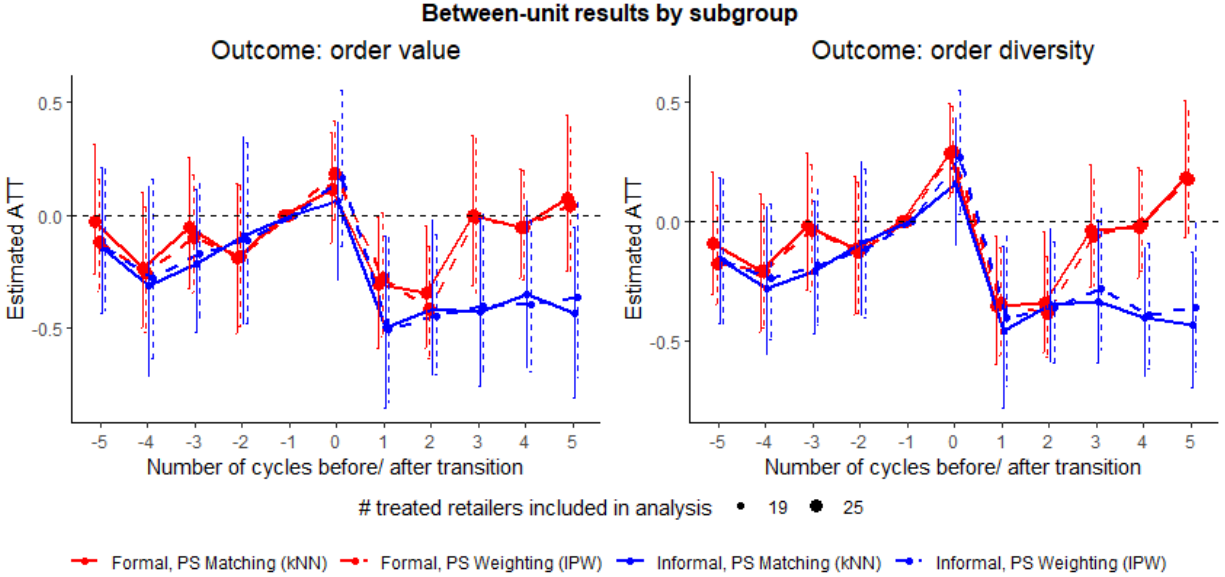


Figure Notes: Solid lines indicate results for propensity-score matching using the k -nearest neighbor algorithm, and dotted lines indicate results using inverse propensity score weighting.

Figure 7: ATT and 95% confidence intervals for formal and informal retailers using the between-unit estimator.

they sold after a sales executive re-assignment, this might result in results similar to the ones we have obtained here. In the appendix, we are able to rule out a change in price of the average product sold, as well as sales of durable (as opposed to consumable) products.

We note that we are unable to reproduce the between-unit results using months as the time unit because of insufficient balance on key covariates for informal retailers. This lends further importance to the use of cycles as our time unit; within the informal subgroup, treated and control retailers are balanced on key covariates within the order cycle preceding treatment, but not necessarily in the month preceding treatment.

2.6 Discussion & Concluding Remarks

The previous section's results highlight the outsized role that personal relationships, and the trust that they embed, play when interfacing with informal retailers – such as retailers serving BoP customers. They have several immediate managerial implications for distribution of life-improving durable goods to BoP consumers and, more generally, commercial settings where

formal governance institutions are weak.

First, our results emphasize the importance of sales executive retention and retailer-relationship management for BoP-focused organizations and distributors. A recent BoP Last-Mile Distributor (LMD) survey found that most LMDs engage with retailers as a sales channel and that most LMDs perceive sales executive retention as a significant challenge (Collective 2019). Our results provide quantitative support for the importance of overcoming this challenge to ensure continued retailer engagement and partnership. Future research may seek to identify the determinants of sales executive attrition, particularly as it varies across sales executive tenure and other characteristics, for the sake of improving retention. The revenue loss associated with a disruption in the sales executive-retailer relationship is further compounded by the additional training costs associated with hiring and training a new sales executive. Since the negative effect of disruption to a sales executive relationship is most salient for informal retailers, which often cater to more vulnerable consumers than formal retailers, such disruptions can significantly hinder an LMD's ability to achieve its social impact goals.

Given these insights, managers at LMDs should aim to proactively prevent the disruption of a relationship between an informal retailer and sales executive once it is in place. Care for this relationship should be reflected in human resources strategies, even when informal retailers or customers seem to be performing worse than expected. In this setting, although the distributor sometimes initiated sales executive reassignments to improve the quality of service provided to a retailer, the consistent failure of informal retailers to recover from the disruptive effect of a transition suggests that other approaches to improving retailer performance may be worthy of consideration. We believe that these findings are generalizable well beyond the setting in which this study was conducted, as informal economy dynamics are similar across emerging markets.

Second, our findings provide empirical evidence for the importance of social relationships in supply chains that cater to informal retailers in the BoP. When a sales executive transition occurs, the products, prices, and financing terms offered by the distributor to the retailer do not change; yet, for informal retailers, order value and diversity are negatively affected for a significant period of time. This result can be interpreted using the sociological concept

of embeddedness (Granovetter 1985), which argues that economic rationality is “embedded” within social ties. The embeddedness literature has found that, on the one hand, social ties facilitate economic activity by creating trust, building economies of time, and allowing for joint problem solving. On the other hand, excessive dependence on social ties makes firms more vulnerable to exogenous shocks and disruptions (Uzzi 1997). Our empirical results are consistent with these findings, and highlight the importance of embeddedness in operations management in the BoP context. For informal retailers, which are less subject to legal and governmental regulations than formal retailers, social ties play a key role in supply chain transactions. While these social ties might increase supply chain efficiency for informal retailers, our data shows that they also amplify the consequences of supply chain disruptions.

These results may partially explain the historical failure of alternative business models (such as traditional and assisted e-commerce) to penetrate BoP markets: these models depend on reducing in-person interaction in order to increase efficiency, which may backfire in markets containing large numbers of informal retailers whose performance is heavily influenced by social relationships.

Given the limitations of our dataset, we could not determine the exact mechanism through which business relationships influence retail channel performance. Future research may examine, perhaps through experimental methods, which aspects of the sales executive relationship can be used as a lever to influence retailer performance. Such work may guide the design of compensatory measures which can mitigate the effects of unavoidable relationship disruptions on informal retailers.

Finally, our results indicate that “classical” supply chain relationship management best practices, such as the development of “deep” long-term suppliers relationships (Liker and Choi 2004), are very valuable in the BoP context. Our partner distributor sells novel life-improving durable goods (such as solar lamps) that are new to both retailers and customers. Thus, an enduring relationship between a sales executive and an informal retailer might lead to the development of joint technical capabilities for marketing and selling these products and a structured lexicon and information exchange process for addressing issues related to product sales. Our own survey data indicates that many informal retailers have low

technological capabilities (see Table 3). In such a setting, the relationship between the sales executive and retailer over time becomes a technology in itself, and this technology is lost once the sales executive leaves. BoP distributors should invest in, learn about, and actively engage informal retailers as an extended part of their organizations.

2.6.1 Alternate Time Units

We now repeat the analysis from Section 2.5.1 using months, rather than cycles, as the relevant time unit. We remove outliers in the same manner as in Section 4.4, but with different thresholds given the different distribution of relative outcomes for months compared to cycles; we use thresholds of $X = 12$ and $X = 6$ for order value and order volume, respectively, for formal retailers; for informal retailers, these values are $X = 7$ and $X = 10$, respectively. After outliers are removed, the sample used in the analysis consists of 34 formal retailers and 28 informal retailers. Results are shown in Figure 8.

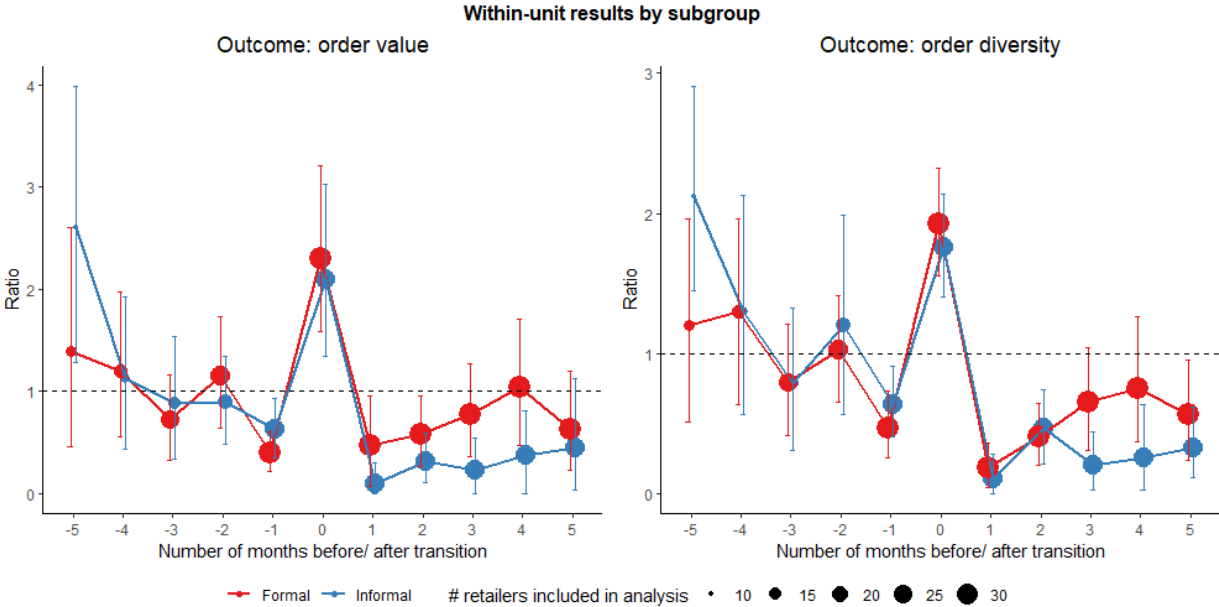


Figure 8: Within-unit ATT with alternate time unit (months)

For order value, formal retailers experience a decrease in performance (relative to pre-treatment means) of 53.3% in period $t_0 + 1$, and informal retailers experience a corresponding reduction of 90.9%. For order diversity, formal and informal retailers experience reductions of 81.6% and 89.3%, respectively, in the first post-treatment period. These results are similar to those obtained with cycles as the time unit, strengthening our overall conclusions: formal retailers recover sooner, while informal retailers do not experience sustained recovery. In this case, the outcomes for formal retailers begin to fall in month five, perhaps reflecting seasonality or time-specific effects which are not accounted for here, but which are controlled

for in the results shown in Section 2.5.2.

Chapter 3

The impact of streak-based performance evaluation in crowdsourcing contests for skilled microtasks

Olumurejiwa A. Fatunde

Center for Transportation and Logistics, Massachusetts Institute of Technology

Joann F. de Zegher

Sloan School of Management, Massachusetts Institute of Technology

3.1 Abstract

Crowdsourcing is used in many settings to organize distributed work, including skilled tasks. Competitive crowdsourcing contests are widely used to elicit skilled knowledge work in software, design, and research. To be successful, crowdsourcing contests need to be designed such that they are effective at eliciting a sufficient number and level of quality outputs from participants. They can do so by recruiting more participants, engaging them in completing a large number of tasks, ensuring sufficient accuracy on tasks, and inducing participants to complete tasks at a decent speed. Using data on 5,418 crowdsourcing contests for medical diagnosis, we examine how these outcomes are impacted by a contest design where partici-

participant performance is evaluated based on *streaks*, *i.e.*, the longest sequence of correct answers, rather than the traditional *accuracy*-based evaluation metric, *i.e.*, % correct answers. We find that streak contests increase the percentage (number) of responses that surpass a quality threshold by 3.5% (11.9%), thereby increasing the number of responses for which there is a consensus on the correct diagnosis. This increase in quality is partly due to the fact that streak contests are particularly effective at driving engagement, with each participant providing 83% more responses in streak contests. Finally, streak contests also increase the speed of achieving consensus by 31%. These results follow from a fundamental difference between accuracy and streak contests; in the latter, randomness and costly recovery from mistakes play a more significant role in determining overall performance, changing participants' incentives. Overall, we find that a streak-based evaluation metric is a viable tool for increasing participant activity and platform performance, particularly in settings where additional participation is not costly, and volume and velocity of activity are important. These findings are relevant to other contexts in which rewards based on performance streaks are used to boost participant motivation; we find that, in addition, they boost performance.

3.2 Introduction and Literature

Recent technological advances have enabled new modes of organizing work, including an increasingly common mode known as crowdsourcing. Crowdsourcing involves organizing a large group of people to perform tasks that a company might otherwise have performed internally. It is often used to solicit innovative software or analytics solutions (*e.g.*, TopCoder, Kaggle), obtain product designs or prototypes (*e.g.*, InnoCentive), maintain accurate mapping resources (*e.g.*, Waze, OpenStreetMaps), contribute to knowledge repositories (*e.g.*, Wikipedia), complete "microtasks" (*e.g.*, Amazon's Mechanical Turk, Google's Crowdsourc) or tag images, audio, or text data (Garcia Martinez and Walton 2014, Wang et al. 2017, Tajedin et al. 2019, Hossain and Kauranen 2015). In this paper, we focus on crowdsourcing *contests*, wherein a firm crowdsources tasks from external participants who compete for a reward based on a pre-defined evaluation metric.

The design of crowdsourcing contests is studied in literature on formal contest theory (*e.g.* see Dechenaux et al. 2014, Segev 2020 for a review) and in literature on the use of

technology to organize human effort (*e.g.* see Bigham et al. 2014 for a review). Recent work has examined the impact of prize structure, instruction design/wording (Jiang et al. 2021b), feedback (Jiang et al. 2021a) and intra-participant knowledge-sharing (Jin et al. 2021) on *participant performance*. Shao et al. (2012) and Chen et al. (2014) explore determinants of contest *participation*, and find that higher prizes and a lower number of competing contests attract more participants. Researchers have also explored the effect of rank feedback (*i.e.*, information about current ranking position) on *participant effort* (Gill et al. 2019) and the presence of highly talented participants (Zhang et al. 2019) on *participant motivation*. Mason and Watts (2010) study the effect of changing the denominator of an accuracy-based evaluation metric – paying participants on a per-puzzle basis rather than per-word basis – on participant effort. We are not aware of other studies that have studied the impact of evaluation metrics in crowdsourcing contests.

Yet, the marketing and behavioral economics literatures have established the role of (“*streaks*”) of activity in influencing participant motivation and effort in sports, games, educational tools, and organizational settings. For example, popular language-learning app Duolingo encourages participants to return every day so as not to lose their streak of activity. Huynh et al. (2018) assessed the impact of this feature and found that streaks keep more advanced participants motivated by making the game more challenging, and improve game attractiveness as a game progresses, with streaks accumulating value at an increasing rate as they grow longer. Levy (1996) suggests that streaks serve as a sort of asset; a combination of loss aversion and endowment effect may lead participants to adopt more risk-averse behavior after accumulating a significant streak. An extensive literature in behavioral economics has established that people tend to under-estimate the likelihood of streaks and thus mis-attribute their performance to other causes such as advanced skill levels or systemic error, leading them to change their behavior in response to those (incorrect) perceived causes (IaElena Asparouhova et al. 2009, Rabin 2002, Andrikogiannopoulou and Papakonstantinou 2018, Gilovich et al. 1985, Hilary and Menzly 2006, Chen et al. 2016, Criscuolo et al. 2021, Kong et al. 2020). While there has been no direct evaluation of performance streaks in the context of crowdsourcing contests, these past examples suggest that crowdsourcing contest participants may under-estimate their ability to continue performing when they are on a

performance streak, thus motivating them to put in more effort to over-compensate for this perception. Hence, performance streaks exhibit many interesting characteristics that may influence participation, engagement and accuracy in work-related contests such as crowdsourcing contests. However, to our knowledge there has been no such analysis of the effect of using performance streaks as an evaluation metric. This is therefore the topic of this paper.

We study the effect of using an evaluation metric based on a *performance streak* (*i.e.* the length of a worker’s longest sequence of correct answers) relative to a more common evaluation metric based on *accuracy* (*i.e.* the % of correct answers), on contest participation, engagement, accuracy, and completion speed. We focus on a canonical contest type called rank-order contests (first studied in-depth by Lazear and Rosen 1981), in which participants are ranked and then awarded a prize based on their rank.

We study this question in the context of a healthcare software company, *CentaurLabs*, that develops crowdsourcing contests for medical diagnosis by creating diagnostic microtasks on its mobile app, *DiagnosUs*. Medical misdiagnosis is ubiquitous, and can have significant (sometimes fatal) consequences for patients. As such, initiatives that improve the accuracy of diagnosis or facilitate the ability of external expertise to give input on diagnostic cases have the potential to transform healthcare. Access to the right expertise to accurately diagnose a problem is in many cases a function of geography. Being able to leverage distant sources of medical knowledge, such as through CentaurLabs’ solution, provides an opportunity to broaden access to quality care.

CentaurLabs uses a “collective intelligence” algorithm to aggregate opinions from the best-performing contest participants and converge on a diagnosis. In each contest, participants are tasked with labeling a sequence of diagnostic microtasks requiring specialized medical knowledge. The contests draw on a mix of problems pre-labeled by experts and unlabeled problems; participant performance on the pre-labeled tasks is used to evaluate skill levels, influencing the weight placed on a participant’s responses to unlabeled tasks. Responses to unlabeled tasks are aggregated to deliver a “consensus label” for each previously unlabeled task.

Participants are workers with different initial skill levels and diverse backgrounds; in a recent survey, 28% of survey respondents indicated they had no formal medical training.

They are motivated by the opportunity to learn, by competition, and by financial compensation for correct answers.

In many crowdsourcing contests, organizers benefit only from submissions of sufficiently high quality. In the context of collective intelligence, the organizers derive value from the aggregated output of contest participants. The goals of crowdsourcing in this context must therefore instead focus, in the short term, on improving the volume and percentage of quality output as efficiently as possible, which is the focus of this paper. Understanding how different performance metrics influence these outcomes provides managerial insight into how crowdsourcing platforms can most effectively leverage distributed suppliers of knowledge and skills for crowdsourcing and other emerging models of organizing work. We find that a metric based on *performance streaks* results in a higher percentage and number of high-quality submissions, as well as a higher velocity of platform activity.

3.3 Data and Key Definitions

The dataset contains crowdsourcing contests run between September 2020 and January 2022. Each observation in the dataset corresponds to a single contest, and includes information about contest parameters, final rankings, and prizes paid. We focus on multiple-choice contests with a single correct answer choice, as this is the group that includes “streak” contests. Table 1 provides a breakdown of contests included in our analyses.

	Number of contests
# Contests starting and ending between 1 Sep 2020 and 31 Jan 2021	5,418
Excluded groups:	
# Contests discarded (e.g, due to setup errors)	395
# Contests run for irrelevant purposes (e.g. testing)	258
# Contests set up for institutions or “squads” of participants*	131
# Contests restricted to a certain audience	1,244
# Contests for which streaks are not relevant**	1,683
# Multiple-choice contests w/ evaluation metrics other than streak/ accuracy	219
# Target contests	1,488

*These contests were excluded due to concerns about network effects and violation of SUTVA assumptions

**e.g., multi-answer, non-multiple choice, “pool” contests

Table 1: Exclusion criteria for contests on the platform

Each *contest* has a pre-defined *topic* (e.g. Knee X-Ray or EEG) and can have multiple

participants. Each participant is assigned a set of *problems* from the *topic* according to a pre-defined but stochastic sampling logic which aims to maximize breadth of coverage across problems and minimize repeat exposure. Thus, for a given contest, each participant will receive a different subset of problems (or, at a minimum, the same problems in a different order), with subsets of problems overlapping across participants.

3.3.1 Contest-level data

Each contest is a time-limited window (lasting between an hour and eleven days) during which a participant solves a set of problems on a given topic. For each contest, we know (1) how problems are scored and winners are determined, (2) restrictions or constraints (*e.g.*, minimum number of problems required to reach the leaderboard or maximum permitted number of problem attempts), (3) "process" information, such as the duration of the contest and the size and structure of the prize basket, and (4) contextual information which is not visible to participants, but may need to be controlled for. Table 3 provides descriptive statistics for each contest-level variable, and Table 2 in the Appendix provides detailed descriptions of these variables.

Contests are competitive; some contests in our data subset have as many as 3,792 participants, and the names of the highest-performing participants are published on a "leaderboard" during each contest. Unlike the typical crowdsourcing contest setting described in the literature (*e.g.*, Zhang et al. (2019)), participants are unaware of the identity of other contest participants before the contest begins. Contests are also incentivized; they offer cash prizes to the most successful labellers, with the total prize bucket per contest ranging between \$1 and \$200 and the number of winners ranging between 1 and 50. For all contests included in this paper, the number of winners as well as the total size and division of the cash pot are announced in advance.

3.3.2 Topic-level data

We have information about the clinical topic for each contest, *e.g.* "Knee X-Ray" or "EEG". We further aggregate topics into larger groups; we consider topics to be part of the same

Variable	n (% of contests)	Mean (St. Dev.)
Evaluation metric = "streak"	322 (21.6)	-
Contest duration (hrs)	-	26.9 (22.88)
Total prize basket (\$)	-	61.86 (32.26)
Minimum prize value (\$)	-	1.06 (2.47)
Maximum prize value (\$)	-	16.52 (8.96)
Maximum number of winners	-	20.1 (8.12)
Evenness/spread of prizes (\$)	-	4.37 (2.73)
Contest has cap on # problem attempts	580 (39)	
Cap on number of problem attempts*	-	241.92 (130.49)
Problems have time limit	355 (23.9)	
Per-question time limit* (seconds)	-	26.63 (9.68)
Contest order within topic	-	16.41 (14.35)
Purpose: customer	1011 (67.9)	
Purpose: engagement	452 (30.4)	
Purpose: research	25 (1.7)	
Audience: cohort	80 (5.4)	
Audience: standard	1408 (94.6)	
Content type: image	1,041 (70)	-
Content type: text	143 (9.6)	-
Content type: video	304 (20.4)	-
Answer choices per problem**	-	3.98 (2.06)
Percentage "free" problems**	-	33 (25)
Platform "crowdedness"	-	6.76 (3.40)

*Average across contests for which restriction is relevant

**For contests with zero participants/responses, values of these variables were imputed based on data from contests within the same topic

Table 2: Descriptive statistics for explanatory variables ($n = 1488$)

"topic group" if they share the same title, prompt, and answer choices. We allow for minor variations in prompt wordings that do not change the meaning of the prompt (*e.g.*, "Do you hear a cough" vs "Is a cough present?"). The 1,488 contest implementations fall within 104 unique "topic groups."

3.3.3 Problem-level data

Each topic consists of a pool of problems. We refer to a problem, or question, as a single piece of medical content (which may come in the form of video, image, or text) for which participants are asked to provide an answer to a multiple choice question. We refer to a response, or a completed problem, as a single instance of a participant providing an answer

to a problem. Participants may occasionally be assigned the same problem more than once. The status of the problem solution (*e.g.*, whether the answer to the problem was provided in advance by an expert, determined via consensus among participants, or is currently unknown) is unknown to participants while they complete each problem. After completing a problem, participants are notified if the problem answer is unknown. The percentage of problems that are pre-labeled ranges from 25% to 100% (mean (50.8%). When user scores are calculated, only their performance on pre-labeled problems is included.

3.3.4 Key decisions

Platform users face a two-stage decision process: first, for a given contest, a user must make a "participation" decision, *i.e.* opting in by beginning the contest. Users can enter contests at any time during the contest window. Secondly, a user must make an "engagement" decision, *i.e.* the number of problems to attempt in the contest. Participants may exit the contest at any time without penalty; their final score and ranking is calculated based on problems completed by the end of the contest.

The following pieces of information are available as inputs to the participation decision:

- *Contest rules & prizes:* Prospective participants are informed of the start and end dates/times of the contest, and the scoring guidelines. The latter includes the minimum participation threshold, the cap on the number of problem attempts, whether the contest is an "accuracy" or "streak" contest, and any tiebreaker rules. This section also indicates how many participants will earn a prize, and the size of the prize for each final rank position. Figures 2 and 3 show how this is communicated visually.
- *Leaderboard:* If the contest has already started, a dynamic "leaderboard" shows the rankings of the top participants, as well as their estimated earnings (*e.g.*, the prize they would win if they maintained their current rank through the end of the contest). The leaderboard also contains links to participant profiles, which contain selected information about participant backgrounds.

Finally, participants can access practice problems on the contest homepage before beginning the contest.

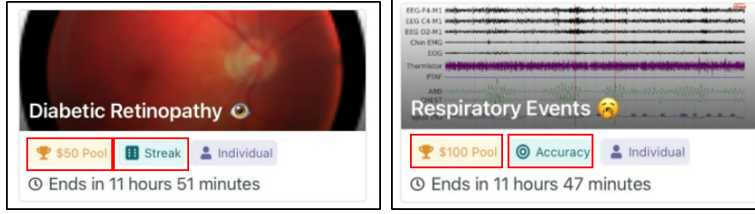


Figure 1: Example of information available to participants on the outer app screen.

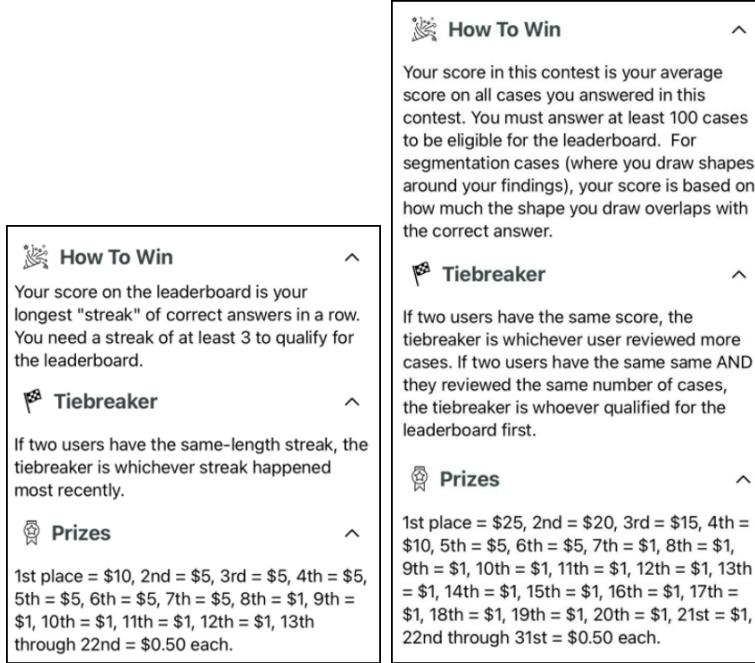


Figure 2: Example information available to participants on scoring guidelines.

3.3.5 Description of contest evaluation metrics

A participant’s performance in a contest can be evaluated using either “streak” or “accuracy” evaluation metrics. Under either metric, participants receive a score out of 100 *for each problem*. Specifically, the score of participant j on problem i in contest k is:

$$\text{score}_{ijk} = \begin{cases} 100 & \text{if participant } j \text{ answers problem } i \text{ in contest } k \text{ correctly} \\ 0 & \text{otherwise.} \end{cases}$$

Under the “accuracy” evaluation metric, the Contest Score (CS) for participant j in

contest k is defined by the average score across all problems in the contest, *i.e.*

$$CS_{jk, \text{accuracy}} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} \text{score}_{ijk}, \quad (3.1)$$

where N_{jk} is the total number of problems completed by participant j in contest k . Participants are ranked based on their values of CS_{jk} , and these rankings determine participant earnings.

Under the "streak" evaluation metric, the contest-level score is defined by the maximum number of consecutive correct answers (a "streak"). The length of a streak at the time of completing problem i reflects the number of consecutive questions that the participant answered correctly most recently:

$$\text{streak}_{ijk} = \begin{cases} \text{streak}_{(i-1)jk} + 1 & \text{if } \text{score}_{ijk} = 100 \\ 0 & \text{otherwise,} \end{cases}$$

and $\text{streak}_{1jk} := 0$. Thus, streak_{ijk} stores the length of participant j 's current streak in contest k after completing problem i ; it is incremented by one for each correct answer and re-set to zero for an incorrect answer. At the end of contest k , the contest-level streak for participant j is given by:

$$CS_{jk, \text{streak}} = \max \{ \text{streak}_{ijk}, \forall i \in [1, N_{jk}] \},$$

i.e. the participant's longest streak achieved at any point during the contest.

Figure 1 shows the median contest score and streak length as a function of the final leaderboard ranking for contests that used an accuracy performance metric *vs.* a streak performance metric, illustrating that the metrics create fundamentally different incentives.

Given that the prize associated with each ranking position is announced in advance, participant j 's expected earning for contest k can be calculated as the expected value of prizes earned, which is a function of the value of each prize and participant j 's probability of winning each prize:

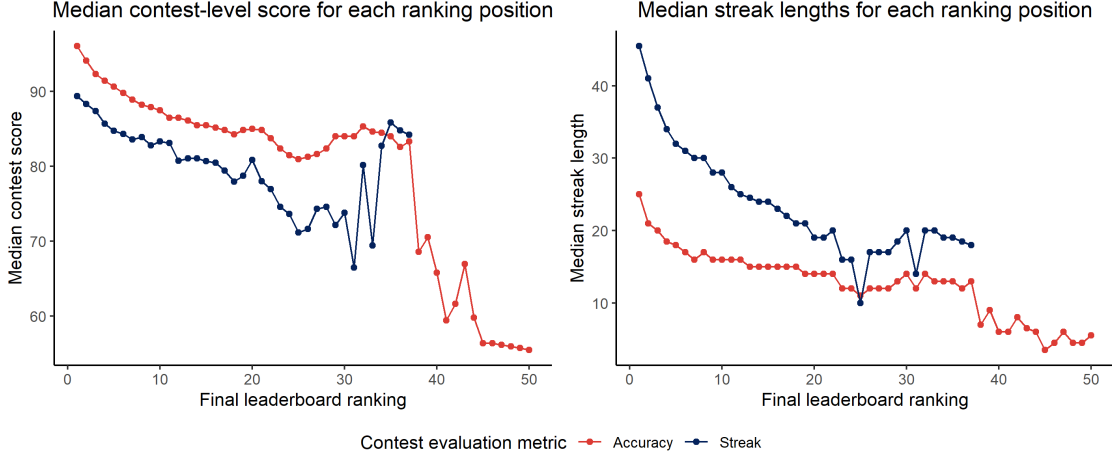


Figure 3: Median contest scores and length of longest success streak for contests using accuracy-based vs. streak-based performance metric

$$\mathbb{E}[earnings_{jk}] = \sum_{\ell=1}^N p_{\ell,j,k} P_{\ell,k}$$

where N is the number of prizes available in contest k , $p_{\ell,j,k}$ is participant j 's probability of attaining a certain final ranking ℓ , and $P_{\ell,k}$ is the prize associated with that final ranking. Any characteristic that increases the value of $p_{\ell,j,k}$ for a given user and contest increases the user's expected earnings, and thus can reasonably be expected to influence user behavior.

Two key differences between accuracy- and streak-based metrics seem particularly relevant to participant decision-making. First, randomness plays a bigger role in determining the final ranking of participants in streak contests. In particular, the random nature of problem allocation order may result in the same participant achieving a different combination of performance streaks (and thus a different value of longest streak length) under different orderings of the same set of problems, despite having the same average score in both cases.

Secondly, the cost of getting a problem wrong, as measured by the amount of effort required for a participant to "recover" and continue improving ranking position after a wrong answer, is higher for a streak contest. In an accuracy contest, users can continue improving their ranking position in one of two ways: 1) they can improve their ranking *passively*, without improving overall score (for example, if a higher-ranked participant j' experiences a reduction in $CS_{j'k,accuracy}$), or 2) they can improve their ranking *actively* by sufficiently

increasing the $CS_{jk,accuracy}$ relative to the contest-level scores of other participants. The number of additional problems required to regain the previous value of $CS_{jk,accuracy}$ after answering a problem incorrectly increases nonlinearly with C_{jk} (the number of problems answered correctly thus far), and can be as low as one (especially for large values of N_{jk}). For streak contests, participant j 's ranking can only be improved *actively* (e.g., by improving the value of $CS_{jk,streak}$, by increasing the length of participant j 's longest streak); the rankings of higher-ranked participants are sticky, since longest streak length remains unchanged even after a participant has lost a streak. Furthermore, the number of additional problems required to regain and extend a streak after losing it increases linearly with the length of the current longest streak. This results in a greater effort being required for recovery in a streak contest.

The greater role of uncertainty and the higher effort required for recovery in streak contests forms the basis of our prediction regarding the effect of streak evaluation metrics on contest participation. This uncertainty may be beneficial to low-skilled participants because it will sometimes result in them winning prizes that are larger than expected. The additional uncertainty inherent in streak contests may also make them less attractive to participants whose decision-making is affected by cognitive biases such as risk aversion (though we do not measure risk aversion in this setting). As such, we expect to see a greater positive effect of streak evaluation metrics on participation in participants with lower subjective or objective skill levels. Although we are unable to directly measure skill level, we measure the impact of streak evaluation metric on the mix of participants using measures such as average experience and average past ranking of participants.

To make a prediction about the impact of a streak evaluation metric on conditional engagement, we consider that overall streak length is proportional to (and bounded by) the number of problems that a participant completes. Therefore, we hypothesize that using a streak evaluation metric will have a positive effect on engagement metrics such as # problems completed on both a contest and participant basis, as participants both attempt more problems (in order to have a chance of extending their streaks) and try harder to get each problem correct (in order to avoid losing their streaks, given the greater cost of recovery).

We expect any effect on performance to come from a change in the mix of participants taking part in contests. Given our prediction of increased relative participation from lower-skilled participants, we expect an overall reduction in performance as measured by the percentage of problems answered correctly.

Finally, we expect that speed of problem completion—and by extension, speed of producing new consensus-based labels—will increase as participants attempt to complete more problems within the contest duration.

3.3.6 Treatment variable

Our treatment variable is the scoring evaluation metric, *i.e.* whether the participant’s rank is based on the length of the participant’s longest performance streak (treatment) or accuracy (control).

3.3.7 Outcome variables of interest

The primary outcome of interest is the number (and percentage) of responses on the platform which meet a certain quality threshold. This quality threshold is a lagging indicator of average performance. Specifically, a response by user j to problem i is considered “qualified” if the trailing average accuracy exceeds a certain threshold (*i.e.*, if the following holds true):

$$\frac{1}{n_{qk}} \sum_{i=n_{qk}+1}^i \text{score}_{ijk} > QT_k$$

where QT_k is the pre-defined quality threshold for a particular contest and n_{qk} is the number of lagging problems over which this accuracy metric is calculated.

The number of qualified responses is therefore defined as the number of responses for which this condition holds, while the percentage of responses qualified compares this number of qualified responses to the total number of responses available.

Because the qualified response metrics consist of multiple components, several secondary outcome metrics can help us understand the underlying reasons for changes in our primary outcome measures. Specifically, there are four key outcome categories at the contest level: (i) participation, (ii) engagement, (iii) performance on labeled problems, and (iv) completion

speed. We consider various ways to measure the outcome variables in each category; Table 4 provides an overview of the metrics used. Some outcome measures in Table 4 rely on a subset of the full dataset; for example, contests which attracted no participants are included in the analysis for participation outcomes, but not in the analysis for outcomes which are conditional on positive participation levels. For each metric, n indicates the number of observations included in the pre-matching dataset (see next section for Methods).

Table 4 describes the relationship between each of the secondary metrics in Table 4 and the primary outcome measures.

Category	Outcome Variable	Metric	n
Participation	Participation	# participants who completed ≥ 1 problem	1488
	Normalized participation	$100 \times \text{Participation} / \# \text{ active in past 15 days}$	1488
Engagement	# problems completed	# responses received from all participants	1477
	Avg. # problems completed / part.	Participant problem count	1477
	# excess responses	Total # responses above min. thresh.	1477
	Avg. # excess responses / part.	Participant responses - min. thresh.	1477
	Median response duration (sec)	Amount of time spent on a problem	1477
Performance	Accuracy	% labeled questions answered correctly	1477
	Avg. # correct responses per partic.	# of questions answered correctly	1477
	# qualified responses	responses meeting a quality threshold*	1477
	Percentage responses qualified	# qualified responses / total # responses	1477
Speed	Net labeling rate**	$(\# \text{ new problems with consensus} - \# \text{ new problems without consensus})/\text{hr}$	1117

*The qualification threshold is a contest-specific minimum level of trailing accuracy

**Includes labels produced between the start time of the current contest and the submission time of the next contest's first response

Table 3: Contest-level outcome variables.

Secondary outcome category	Influence on primary outcome(s)?
Participation	Increased participation increases the density of participation per problem, influencing engagement metrics
Engagement	Increased engagement increases both the number of qualified responses and the number of total responses. Overall effect depends on accuracy
Performance	Increased accuracy increases the number of qualified responses
Speed	Positively with # qualified responses, as a greater number of qualified responses increase the likelihood of achieving consensus

Table 4: Relationship between secondary outcomes and primary outcomes.

Table Notes: *(+) indicates an improvement relative to the control mean, while (-) indicates decline.
 **Three stars indicate significance at $\alpha = 0.01$, two stars indicate significance at $\alpha = 0.05$, and one star indicates significance $\alpha = 0.1$

3.4 Methodology

3.4.1 Causal effect of interest

Our goal is to measure the causal effect of a streak-based evaluation metric on contest-level outcome variables of interest. The desired causal effect is the Average Treatment Effect on the Treated (ATT), given by:

$$\hat{\tau} := \mathbb{E}[Y_k(1) - \hat{Y}_k(0) | T_k = 1], \quad (3.2)$$

where T_k is the treatment indicator for contest k , $Y_k(1)$ is the outcome of contest k under treatment and $\hat{Y}_k(0)$ is the estimated outcome under control. While $Y_k(1)$ is observed for all treated contests, $\hat{Y}_k(0)$ is never observed, and thus must be imputed. We use matching methods to identify counterfactual observations that allow us to impute these potential outcomes under control, and then use a simulation and bootstrapping procedure to impute the potential outcomes and obtain the ATT, given by $ATT_{o,\text{sim}}$ for outcome o :

$$ATT_{o,\text{sim}} = \frac{1}{|n_1|} \sum_{k=1}^{|n_1|} \left[Y_{o,k} - \mathbb{E}[Y_{o,k}(T_k = 0)] \Big| T_k = 1 \right],$$

wherein n_1 is the set of treated contests, and $|n_1|$ is the number of contests in this set. (Similarly, n_0 and $|n_0|$ refer to the set of control contests and the number of control contests, respectively).

In order to identify the causal effect, we need the following assumptions to hold: SUTVA (no interference), common support (nonzero possibility of each treatment option), and conditional ignorability. We believe it is reasonable to expect that the SUTVA assumption holds, *i.e.* that the vast majority of participants compete in contests independently. This is because (i) we have eliminated contests which explicitly rely on relationships between participants (*e.g.*, squad contests, where it would be unreasonable to claim that decisions among participants are independent), and (ii) *CentaurLabs* takes a number of steps to prevent cheating (*e.g.*, in the form of the same participant having accounts on multiple devices). To satisfy the common support assumption, we exclude topic groups for which all contests are either

only streak-based contests or only accuracy-based contests. Conditional ignorability, or the independence of treatment and outcomes conditional on explanatory variables, is addressed through matching.

3.4.2 Matching

We use matching methods to ensure that we compare contests that are similar to one another on all dimensions other than the treatment variable. We combine exact matching (including Coarsened Exact Matching on certain variables), caliper restrictions, and nearest-neighbor matching.

We implement our matching procedure via the MatchIt package in R (Ho et al. 2011). We must implement the procedure separately for each outcome because some of the descriptive variables are only relevant for some of the outcomes (*e.g.* data on contests with no entries can only be used for outcomes related to participation). In this section, we describe the matching procedure for the outcomes which use the full dataset ($n = 1,488$). Summaries of the matching process for outcomes that use smaller data subsets (as described in Table 4) are available in the Appendix.

We require exact matching on variables that significantly alter the nature of the contest: the topic group, the targeted audience (see Table 2 for a definition), and whether there is a cap on the maximum number of problems that a participant can attempt in a contest.

We also require exact matching on a discretized version of minimum prize value, given the extreme lumpiness in the distribution of this variable, for which most contests take on a value of \$0.5 or \$1 (see Figure 4). This has the effect of “coarsening” the minimum prize value variable; known as Coarsened Exact Matching (CEM), and avoiding a level of restrictiveness that might be too high to be useful. CEM has been shown to be an effective way to modify other matching methods in order to reduce bias and achieve balance (*e.g.* Austin 2014, Austin and Stuart 2017a,b).

To maximize balance, we further use a caliper to restrict matches; for a control contest to be matched to a treated contest, we require that its basket size fall within \$10 of the one of the treated contest, and that the number of winners and the number of other contests running at the same time be no more than 1 greater or less than the corresponding value for

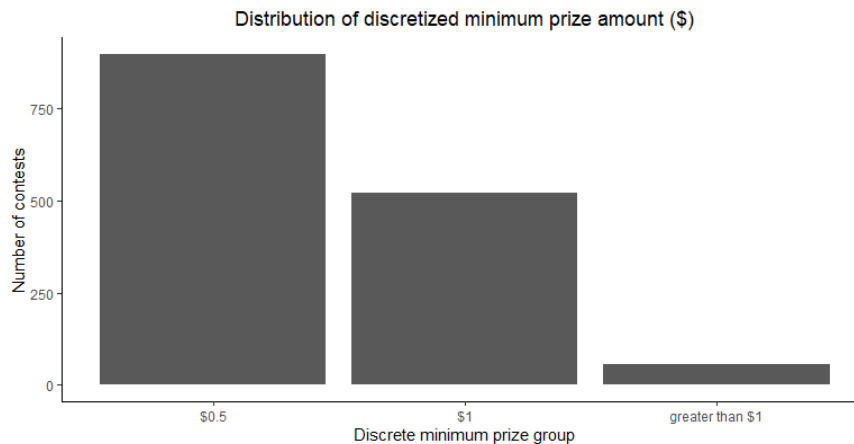


Figure 4: Discretized version of minimum prize value

the treated contest.

Finally, we use k -nearest-neighbor matching to select the set of counterfactual contests for each treated contest. We require that nearest neighbors be chosen from the strata created by exact matching and Coarsened Exact Matching, as described above. Thus, our method combines exact, coarsened, and nearest-neighbor matching, capturing the benefits of each.

We conduct k -nearest-neighbor matching iteratively with different values of k , and select $k = 8$ as the value that resulted in the best balance. Thus, for each treated contest we select as many as eight contests that are most similar to the treated contest, using a propensity score that captures the probability of treatment. Propensity scores are calculated via logistic regression, where the outcome measure is the binary treatment variable, indicating a streak or accuracy contest.

Table 5 describes how the dataset is trimmed during the matching process. Because some treated observations were unmatched due to restrictions of the matching process, our sample population is modified, and our estimated ATT thus applies only to this reduced sample.

Table 7 provides an overview of the post-matching balance on covariates in treatment and control contests. To measure balance, we use the standardized mean difference (SMD) as defined in Ho et al. (2011). Specifically, the SMD for covariate t is given by:

$$\text{SMD}_t = \frac{\frac{1}{|n_1|} \sum_{k \in n_1} x_{k,t} - \frac{1}{|n_0|} \sum_{k \in n_0} x_{k,t}}{s_{\text{treated},t}}$$

	# Contests		
	Streak (treatment)	Accuracy (control)	Total
All topic groups (104 topics)	322	1,166	1,488
Discarded due to lack of common support	NA	455	455
Discarded due to exact, caliper, or kNN matching	87	378	465
# contests included in final sample	235	333	568

Table 5: Summary of matching process data sample for outcomes using the complete dataset

for the pre-matching dataset and

$$\text{SMD}_t^m = \frac{\frac{1}{|n_1|} \sum_{k \in n_1} w_k x_{k,t} - \frac{1}{|n_0|} \sum_{k \in n_0} w_k x_{k,t}}{s_{\text{treated},t}}$$

after matching is complete. As before, n_0 and n_1 represent the sets of control and treated contests, respectively, and $|n_0|$ and $|n_1|$ represent the number of contests in each set. $s_{\text{treated},t}$ denotes the (pre-matching) sample standard deviation for all treated units. The matching weight on each observation, w_k , is equal to 1 for treated contests for which a match is found, equal to 0 for unmatched treated or control contests, and equal to values that satisfy the following two conditions for control contests that are matched:

$$w_k \propto \sum_{m \in n_1} \mathbf{1}\{k \in M_m, \}$$

where M_m is the set of matched control contests for treated contest m , and

$$\sum_{k \in n_0} w_k = \sum_{k \in n_0} \mathbf{1} \left\{ \left[\sum_{m \in n_1} \mathbf{1}\{k \in M_m\} \right] > 0 \right\}.$$

The first condition requires that the weight of each matched control contest is proportional to the number of treated contests to which it was matched. The second condition requires that the sum of all control weights equals the number of control contests that were matched to at least one treated contest.

An SMD below 0.1 is commonly considered an acceptable level of imbalance between

groups (Wang et al. 2013). Table 2 in the Appendix provides definitions for all covariates.

Variable		Before matching			After matching		
		μ_T	μ_C	SMD_t	μ_T	μ_C	SMD_t^m
Format	Content type: image	0.75	0.68	0.16	0.72	0.72	0.00
	Content type: text	0.12	0.09	0.1	0.12	0.12	0.00
	Content type: video	0.12	0.23	-0.31	0.16	0.16	0.00
	Answer choices per problem	3.76	4.04	-0.15	3.70	3.72	-0.00
Process	Contest duration (hrs)	36.75	24.18	0.39	34.27	32.58	0.05
	Minimum prize value: \$0.5	0.63	0.61	0.05	0.70	0.70	-0.00
	Minimum prize value: \$1	0.24	0.38	-0.33	0.23	0.23	0.00
	Minimum prize value: > \$1	0.13	0.01	0.35	0.07	0.07	-0.00
	Maximum prize value	14.57	17.06	-0.23	15.00	15.26	-0.02
	Evenness/spread of prizes	3.64	4.57	-0.28	4.00	4.15	-0.05
	Maximum number of winners	17.55	20.80	-0.40	18.48	18.39	0.01
	Total prize basket (\$)	48.87	65.44	-0.52	53.77	52.92	0.03
Restrictions	Cap on # problem attempts*	7679	5786	0.46	7152	7150	0.00
	Per-question time limit (s)**	3587	4847	-0.43	3483	3422	0.02
Environment	Contest order	13.17	17.3	-0.36	15.37	14.68	0.06
	Purpose: customer	0.39	0.76	-0.75	0.43	0.43	-0.00
	Purpose: engagement	0.61	0.22	0.8	0.57	0.57	0.00
	Purpose: research	0.00	0.02	-0.17	0.00	0.00	0.00
	Audience: cohort	0.07	0.05	0.10	0.02	0.02	0.00
	Audience: standard	0.93	0.95	-0.10	0.98	0.98	0.00
	Platform "crowdedness"	6.73	6.77	-0.01	6.06	6.13	-0.02
Percentage "free" problems	0.24	0.36	-0.49	0.25	0.27	-0.07	

Table 6: Pre/Post-matching balance summary for contest covariates ($n = 1,488$)

Table notes: μ_T and μ_C are the mean across treatment contests and control contests respectively.

*Contests with no cap are coded with a large value (10,000) for matching purposes only. **Contests with no per-question time limit are coded with a large value (6,000) for matching purposes only

3.4.3 Effect estimation

Finally, we estimate the treatment effect. We do this by using standard regression methods, using the trimmed dataset and weights from the matching process. Since the outcome variables have a range of statistical distributions (see Table 8), we select the most appropriate distribution for each outcome using theoretical considerations; *e.g.*, negative binomial distribution to model non-negative count outcomes. For outcomes that might reasonably be described by more than one distribution, we use goodness-of-fit diagnostics to select the most appropriate distribution.

Category	Outcome Variable	Outcome distribution
Participation	Participation	Negative binomial
	Normalized participation*	Gamma
	Avg. experience of participants (days)	Gamma
	Avg. past ranking of participants (%)*	Gamma
	Avg. topic-specific experience (problems)	Normal
Engagement	# problems completed	Negative binomial
	Avg. # problems completed per participant	Negative binomial
	# excess responses	Negative binomial
	Avg. # excess responses per participant	Normal
	Median response duration (sec)	Gamma
Performance	Accuracy	Normal
	Avg. # correct responses per participant	Negative binomial
	% responses qualified*	Gamma
	# responses qualified*	Negative binomial
Speed	Net labeling rate	Normal

*Transformed from percentage to rate over 100

Table 7: Statistical distributions used to model each outcome variable

We run the regression for each outcome and recover the coefficient on the contest evaluation treatment variable. The regression for outcome o is given by the following general form:

$$Y_{o,k} \sim f_o(\theta_{o,k}, \sigma_o),$$

where $Y_{o,k}$ is the value of outcome measure o for contest k , $f_o(\theta_{o,k}, \sigma_o)$ is the probability density function with $\theta_{o,k}$ as the systematic component and σ_o as the ancillary parameter (*e.g.*, variance in the case of a Normal distribution, scale parameter in the case of a Gamma distribution). The systematic component $\theta_{o,k}$ can be further expanded as follows:

$$\theta_{o,k} = g_o(X_k, \beta_o),$$

where X_k denotes a $1 \times n$ vector of data for contest k that includes the intercept term, the treatment indicator, and $n - 2$ contest-level covariates (described in Table 2). Similarly, β_o is a $n \times 1$ vector containing coefficients for each of the covariates in X_k . Specifically:

$$X_k = [1 \quad T_k \quad X_{1,k} \quad X_{2,k} \quad \dots \quad X_{(n-2),k}],$$

$$\text{and } \beta_o = \begin{bmatrix} \beta_{o,0} \\ \beta_{o,1} \\ \beta_{o,2} \\ \dots \\ \beta_{o,n} \end{bmatrix}.$$

When OLS is used to model outcome o , the regression reduces to:

$$Y_{o,k} = X_k \beta_o + \epsilon_{o,k},$$

where $\epsilon_{o,k} \sim N(0, \sigma_o)$ and $ATT_{o,\text{sim}}$ is $\beta_{o,1}$, the coefficient on the treatment indicator. However, for several distributions (*e.g.*, Negative Binomial and Gamma), the regression coefficients are not directly interpretable in a way that translates into treatment effects. We thus use the counterfactual simulation method proposed by King et al. (2000) and Imai et al. (2008a) to generate the ATTs as well as uncertainty measures in a way that is consistent across all outcome measures. We implement this procedure using the Zelig package in R (Choirat et al. 2020).

To conduct the simulation, we first extract the vector of estimated coefficients, $\hat{\beta}_o$, and the estimated ancillary parameter, $\hat{\sigma}_o$, and define $\hat{\delta} := [\hat{\beta}_o, \hat{\sigma}_o]$. We also extract the variance-covariance matrix of the estimates, $V(\hat{\delta})$.

For each draw of the simulation, we randomly sample a parameter vector $\tilde{\delta} = [\tilde{\beta}_o, \tilde{\sigma}_o]$ from the multivariate normal distribution with mean and variance equal to $\hat{\delta}$ and $V(\hat{\delta})$. To calculate the systematic portion of the regression model, $\theta_{o,k} = g_o(X_k, \beta_o)$, we use the sampled coefficients $\tilde{\beta}_o$ and set each covariate to a representative value for each contest k . Specifically, let X'_k represent the vector of representative values for contest k . The element of X'_k corresponding to the treatment variable, T_k , is set to 0 or 1, depending on whether we want to generate expected values of the outcome variable under treatment or control. For all other explanatory variables, the elements of X'_k are set to the observed values of contest k . We can then compute the simulated $\tilde{\theta}_{o,k}$ using the transformation function g_o , the representative values of each covariate for each contest (X'_k), and the simulated coefficients

$\tilde{\beta}_o$:

$$\tilde{\theta}_{o,k} = g_o(X'_k, \tilde{\beta}_o).$$

This procedure allows us to capture estimation uncertainty, caused by having a limited sample size. To capture fundamental uncertainty that may affect the outcome value but that is not included in the systematic component, we also simulate the stochastic portion of the regression model. We take 1,000 draws of the simulated outcome variable, $\tilde{Y}_{o,k}^{(n)}$ (with $n = 1, \dots, 1000$), from the distribution of outcomes using the sampled parameter vector $\tilde{\delta}$, *i.e.* we take 1,000 draws from

$$f(\tilde{\theta}_{o,k}, \tilde{\sigma}_o).$$

Finally, we average over the fundamental uncertainty by calculating the mean of the simulations to obtain a simulated expected value:

$$\tilde{\mathbb{E}}(Y_o) = \sum_{n=1}^{1,000} \tilde{Y}_{o,k}^{(n)} / 1,000.$$

Observe that in the case where $f(\cdot, \cdot)$ is a normal distribution and we use OLS, we have that $\tilde{\mathbb{E}}(Y_o) = g_o(X'_k, \tilde{\beta}_o)$. Thus, the usefulness of the latter step is mainly to accommodate fundamental uncertainty in other, non-linear models. We use $\tilde{\mathbb{E}}(Y_o^{T_k=1})$ and $\tilde{\mathbb{E}}(Y_o^{T_k=0})$ to denote the simulated expected value calculated with T_k set to 1 or 0 respectively for all k .

We repeat this procedure, starting with the generation of $\tilde{\delta}$ and ending with a simulated expected value, 1,000 times. To do so, we use a nonparametric bootstrap procedure, resampling the data with replacement in each repetition. The resulting output serves as the distribution for the expected value of the outcome variable of interest.

The procedure allows us to compute the ATT, defined as follows:

$$ATT_{o,\text{sim}} = \frac{1}{|n_1|} \sum_{k \in n_1} \left[Y_{o,k} - \tilde{\mathbb{E}}(Y_o^{T_k=0}) | T_k = 1 \right].$$

In summary, we calculate the counterfactual term of $ATT_{o,\text{sim}}$ by using the entire matched dataset to generate simulated coefficients ($\tilde{\delta}$), and then generating a simulated counterfactual expected value $\tilde{\mathbb{E}}(Y_o^{T_k=0})$ for treated units with the value of the treatment

variable set to 0 and the value of explanatory variables set to their observed values. $ATT_{o,\text{sim}}$ is calculated as the expectation of the difference between the mean observed value of treated observations and the simulated counterfactual expected value (see also Ho et al. 2011, Imai et al. 2012, Choirat et al. 2020, Imbens 2004). We calculate confidence intervals by taking the relevant percentiles from the sampling distribution of the simulated ATT value.

3.4.4 Heterogeneous Treatment Effects

We expect the treatment to impact different types of contests differently. We thus implement a heterogeneity analysis by computing contest-level treatment effects, which we aggregate to obtain a conditional version of the ATT for subgroups of interest.

To do so, we leverage a class of machine-learning models called “meta-algorithms”, or “meta-learners”, which build on standard supervised learning methods to estimate heterogeneous treatment effects. Specifically, we work with the X-Learner algorithm developed by Künzel et al. (2019) to compute treatment effects at the level of individual contests, and then aggregate them to produce subgroup-level treatment effects. Compared to other meta-learners, the X-Learner is quite flexible, since it allows modeling treatment and control observations separately. This algorithm is thus well-suited to the lopsided nature of treatment in our data. To implement the X-Learner algorithm, we proceed with the following steps in the post-matching sample.

We begin by estimating a response function, $f_o(\theta_{o,k}, \sigma_o)$, separately for control and treated observations, using the distributions described in Table 8 and the same covariates as in the main regression. The output is a set of estimated coefficients for each response equation, denoted by $\hat{\beta}_o^{T_k=1}$ for treated contests and $\hat{\beta}_o^{T_k=0}$ for control contests. To obtain the predicted outcome of a control contest under treatment, we then compute

$$\hat{Y}_{o,k}^1 = g_o(X'_k, \hat{\beta}_o^{T_k=1}) \text{ for } k \in n_0.$$

Similarly, to obtain the predicted outcome of a treated contest under control, we compute

$$\hat{Y}_{o,k}^0 = g_o(X'_k, \hat{\beta}_o^{T_k=0}) \text{ for } k \in n_1.$$

We use these imputed predicted outcomes together with the observed outcomes to compute an imputed treatment effect $\tilde{\tau}_{o,k}$ for each contest k :

$$\begin{aligned}\tilde{\tau}_{o,k}^0 &= \hat{Y}_{o,k}^1 - Y_{o,k} & \text{for all } k \in n_0, \\ \tilde{\tau}_{o,k}^1 &= Y_{o,k} - \hat{Y}_{o,k}^0 & \text{for all } k \in n_1.\end{aligned}$$

After generating the imputed treatment effects for each contest, we use OLS to model the estimated treatment effect, using the imputed individual treatment effects as our dependent variables and the contest-level covariates as the independent variables:

$$\begin{aligned}\tilde{\tau}_{o,k}^0 &= X_k \beta_o + \epsilon_{o,k}^0 & \text{for } k \in n_0, \\ \tilde{\tau}_{o,k}^1 &= X_k \beta_o + \epsilon_{o,k}^1 & \text{for } k \in n_1.\end{aligned}$$

This allows us to generate predicted values ($\hat{\tau}_{o,k}^0$ and $\hat{\tau}_{o,k}^1$) of the treatment effect for each contest.

We then aggregate these to obtain a conditional version of the ATT for subgroups of interest, *i.e.* the Conditional Average Treatment Effect (CATE). We use $\hat{\tau}_{o,X}$ to denote the CATE for a subgroup X ; $\hat{\tau}_{o,X}$ can be computed by aggregating the contest-level predicted treatment effects under treatment and control using a weighted average, where the weights are the estimated propensity scores ($\hat{\pi}_k$) for each contest (see 3.4.2):

$$\hat{\tau}_{o,X} = \hat{\pi}_k \hat{\tau}_{o,k}^0 + (1 - \hat{\pi}_k) \hat{\tau}_{o,k}^1.$$

When reporting results, we compute the CATE only for treated contests, thereby generating a conditional version of the Average Treatment Effect on the Treated (CATT) (Imbens 2004), which we compare to our main results. We only include contests for which $0 < \hat{\pi}_k < 1$, thus mirroring the common support requirement that we used for computing the main effects. We display the computed CATT alongside the 2.5th and 97.5th percentiles of the distribution of contest-level individual treatment effects for additional context.

We examine effect heterogeneity along nine dimensions of the outcome measures, defined in Table 8. For each dimension, the CATT was computed for subgroups selected based on

natural divisions of each variable. Continuous variables were discretized for this purpose.

Heterogeneity dimension	Subgroups
Cap on # of problems	Binary (Yes/No)
Time limit	Binary (Yes/No)
Duration	< 12hrs/12hrs/12-24hrs/24hrs/24+hrs
Max # prizes available	1-10,11-20,21-30,31+
Min prize value	\$0.5,\$1,\$1+
Max prize value	< \$10,\$10,\$10+
Total prize basket value (\$)	< 30,31-75,76-100,100+
# answer choices	2/2+
# alternative contests available	0-5/6-10/11+

Table 8: Subgroups for heterogeneity analysis

3.5 Results

In this section, we show the effect of changing from an accuracy-based to a streak-based evaluation metric on the outcomes of interest.

3.5.1 Primary outcome metrics

Because the quality threshold determines which participant responses are incorporated into the final consensus labels, improvements on “percentage of responses qualified” are of particular interest. The coefficients for the treatment variable are shown in Table 9; Figure 5 shows the estimated ATT for each outcome. We find that streak contests lead to a 2.5 percentage-point increase in the percentage of responses qualified (3.5% increase from a control mean of 69.4%, as well as increase of 1,654 in the *number* of qualified responses (11.9% improvement compared to a control mean of 13,844).

Both primary outcome metrics, percentage and number of qualified responses, increased significantly. In the sections that follow, we will see that these increases are driven by an increase in the number of total responses as well as the number of correct responses.

Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
% responses qualified	-0.0003	0.0001	0.02
# responses qualified	0.41	0.03	<0.001

Table 9: Coefficients, standard errors and *p*-values for the treatment variable for primary outcomes

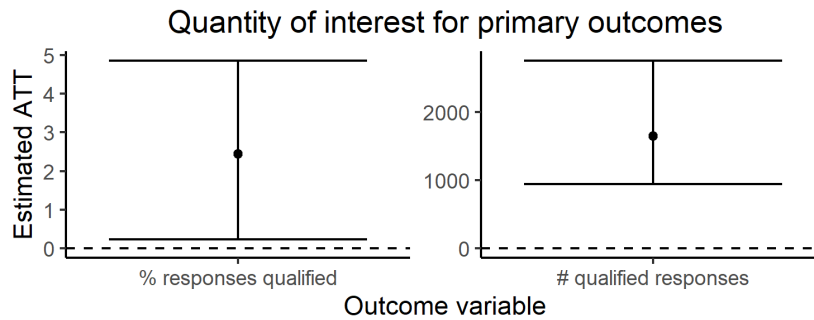


Figure 5: Simulated expected value of treatment effect for primary outcomes

3.5.2 Secondary outcome metrics

Participation

The treatment coefficients that were estimated the regression equations for the two participation outcomes are shown in Table 10. Recall that the distributions for the different outcomes do not always permit a direct interpretation of the coefficients; hence, Figure 6 shows the estimated ATT for each outcome, with standard errors calculated via the simulation method described in Section 4.4.

We find there is no significant impact on the *number* of participants.

Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
Participation	-0.002	0.02	0.93
Normalized participation	0.0007	0.005	0.88

Table 10: Coefficients, standard errors and *p*-values for the treatment variable for participation outcomes

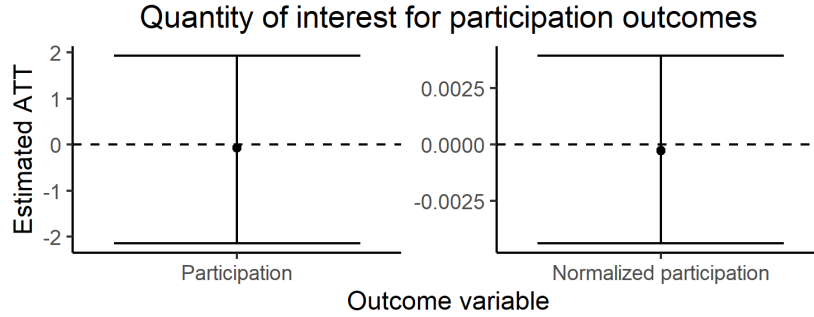


Figure 6: Simulated expected value of treatment effect for participation outcomes

Engagement

The next category of outcomes focuses on participant engagement, given participation. The coefficients for the treatment variable are shown in Table 11; Figure 8 shows the estimated ATTs.

A streak-based evaluation metric has a statistically significant positive influence on all three engagement metrics: it increases the total number of problems completed in a contest by 3,528 (13.9% compared to a control mean value of 25,296) and the total number of problems by participant by 88 (83.3% increase). The change in number of responses completed per participant above the minimum contest threshold is even more striking; this metric increases by 18123 at the contest level (133.5% increase relative to control mean of 13,577) and by 146 (947% increase compared to a control mean value of 15.4 responses above the threshold). Streak contests also reduce the median time spent per problem by 0.025 seconds (compared to a control mean value of 2.15 seconds).

These results are in line with expectations. When using a streak-based evaluation metric, the number of problems completed bound the length of a participant's streak; hence, completing more problems increase a participant's chances of obtaining a longer streak. The higher total number of problems completed is in fact driven by an increase in the number of responses per participant rather than an increase in the number of participants (per the previous section). Because we also find that the number of responses above the minimum threshold increases, we know that participants are completing more problems in order to compete, rather than simply to be eligible for the leaderboard. The average increase in

number of problems completed is 88 for all participants and 361 for top performers, which corresponds to approximately 83.3% and 263.1% more problems completed for each group, respectively. We also find that there is a decrease in the time spent completing each problem, perhaps reflecting participants' desire to complete a larger number of problems. These engagement metrics suggest an increase in the denominator of the % responses qualified, as hypothesized in Table 4.

	Outcome	Coeff.	s.e.	p
	# problems completed	0.39	0.02	<0.001
Avg. # problems completed per participant		0.38	0.02	<0.001
	# excess responses	1.45	0.03	<0.001
Avg. # excess responses per participant		145.8	3.14	<0.001
	Median response duration (sec)	0.011	0.003	<0.001

Table 11: Coefficients, standard errors and p -values for the treatment variable for engagement outcomes

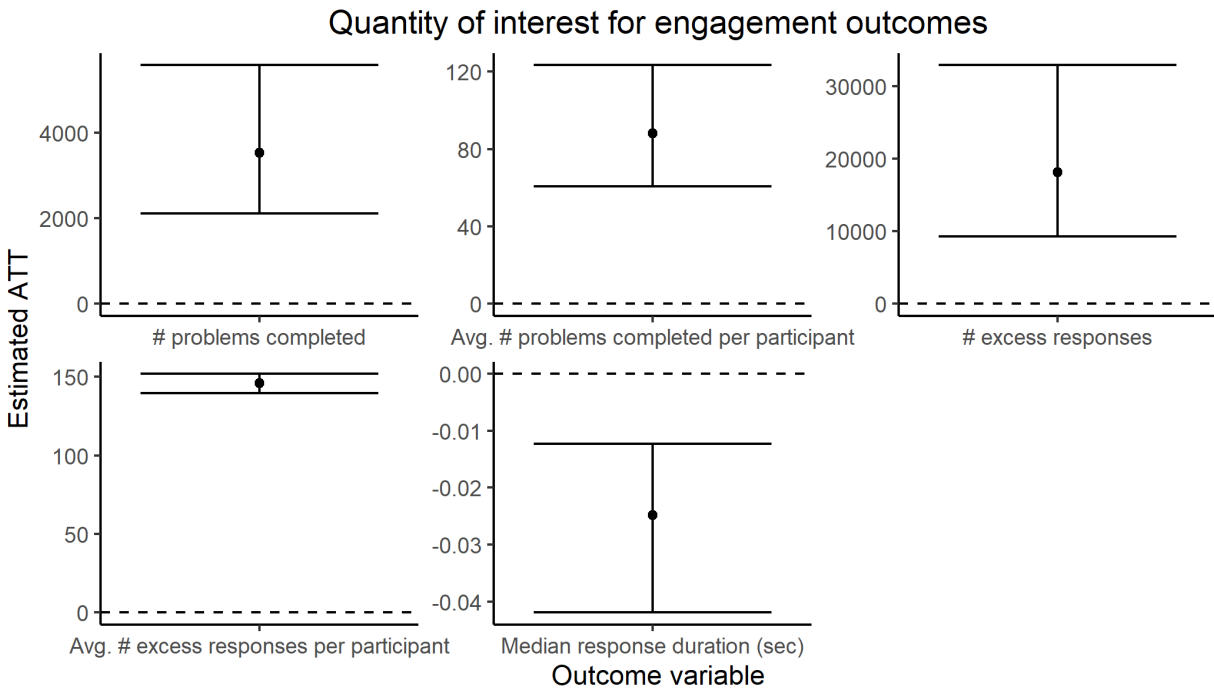


Figure 7: Simulated expected value of treatment effect for engagement outcomes

Performance

The secondary performance outcomes of interest are the number of questions answered correctly for each participant and the contest-level average of participants' problem scores. The coefficients for the treatment variable are shown in Table 12; Figure 8 shows the estimated ATT for each outcome.

Streak contests significantly increase the number of correct responses, which also translates into an increase in the magnitude of the numerator of one of our primary metrics, % responses qualified. However, since the increase in the average number of correct responses (54 additional correct responses per participant and 180 for prize winners) is lower than the increase in the number of problems completed (see previous section), streak contests see a reduction of 1.4 percentage points in the average percentage of problems answered correctly for each participant (4.99 percentage-point reduction for prize winners).

	Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
	Accuracy	-1.42	0.29	<0.001
	Avg. # correct responses per participant	0.42	0.02	<0.001

Table 12: Coefficients, standard errors and *p*-values for the treatment variable for performance outcomes

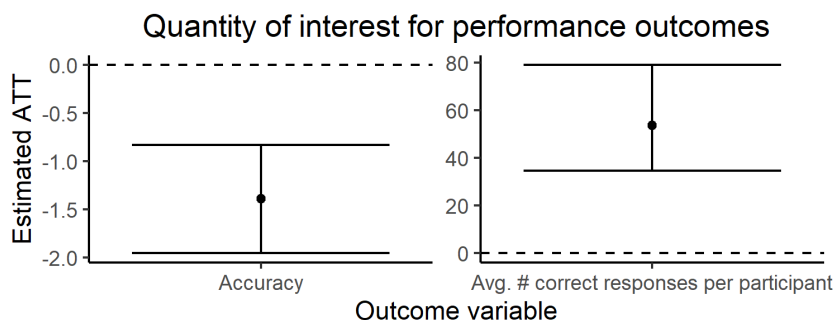


Figure 8: Simulated expected value of treatment effect for performance outcomes

Speed

Our results for labeling rate are obtained using a subset of the data; because labels are tracked at the problem level, we include only those contests which were not run simultaneously with

another contest drawing from the same problem pool. The coefficients for the treatment variable are shown in Table 13; Figure 9 shows the estimated ATT for each outcome.

Streak contests deliver consensus labels at a faster rate; the increase in labeling rate (circa 3 labels per hour) represents an increase of about 31% relative to the average control contest.

Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
Net labeling rate	3.37	1.50	0.03

Table 13: Coefficients, standard errors and *p*-values for the treatment variable for speed outcomes

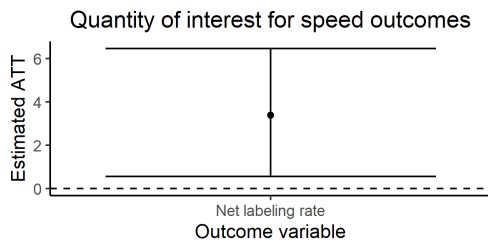


Figure 9: Simulated expected value of treatment effect for speed outcomes

3.5.3 Effect heterogeneity

The X-learner procedure described in Section 4.4 provides insight into the source of the observed results. We find that the most prominent sources of effect heterogeneity are related to the prize basket; (*i.e.*, minimum prize value and total prize basket value, and number of prizes available). Table 14 provides an overview of the dimensions on which we find heterogeneity in the effect of a streak-based evaluation metric. We focus this section on presenting those results.

Category	Outcome Variable	Heterogeneity dimension		
		Min. prize (\$)	# Prizes	Prize basket (\$)
Participation	Participation	✓		
	Normalized participation			
Engagement	# problems completed	✓		
	# problems completed / part.	✓		
	# excess responses	✓		✓
	Avg. # excess responses / part.			
	Median response duration			
Performance	Accuracy	✓		✓
	Avg. # correct responses	✓	✓	
	% responses qualified	✓		✓

Table 14: Summary of effect heterogeneity.

Table Notes: checkmark denotes that the treatment effect varies along the corresponding heterogeneity dimension. We do not examine heterogeneity for net labeling rate due to small sample size in some categories of certain heterogeneity dimensions

Figure 10 illustrates that minimum prize value is a source of effect heterogeneity for several metrics of participation, engagement, and performance. For all metrics where such heterogeneity is present, the effect is most prominent in contests with high minimum prizes. In the case of participation, we previously found that the effect of a streak evaluation metric on the number of participants was null; here we find that there is a significant *positive* increase in number of participants under a streak-based evaluation metric for contests with the largest minimum prize values (\$1 and above). The effect on percentage of responses qualified, while positive overall, is reversed for high-minimum contests. This likely reflects the large increase in additional activity seen by these contests.

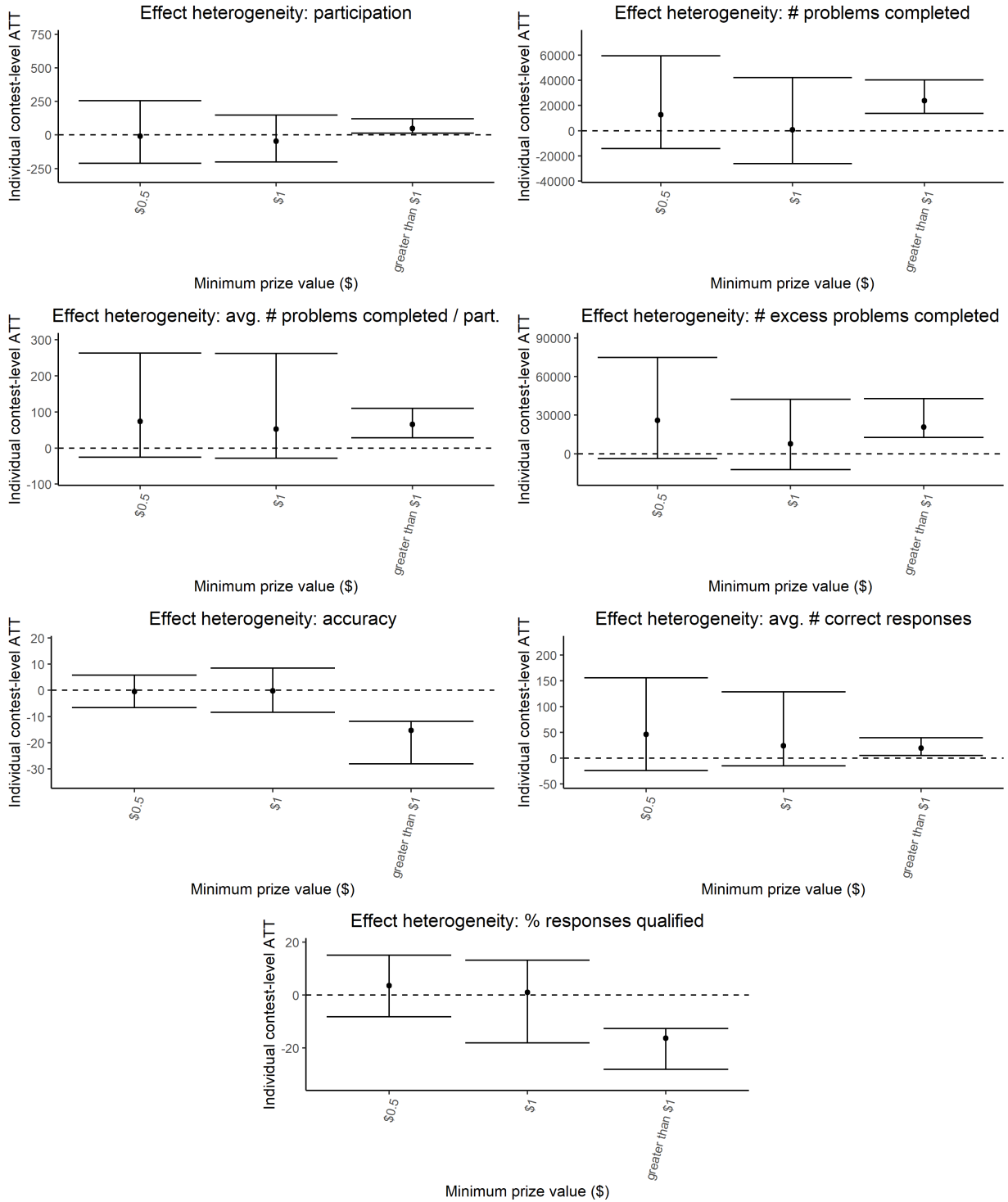


Figure 10: Conditional ATT based on minimum prize value

Figure Notes: Point represents Conditional ATT, and lower/upper bounds represent the 2.5th/97.5th percentiles of contest-level individual treatment effects

Figure 11 shows how the number of available prizes the impact of a streak-based eval-

uation metric on the number of correct responses. This effect, while positive, is strongest in contests with the largest number of prizes available.

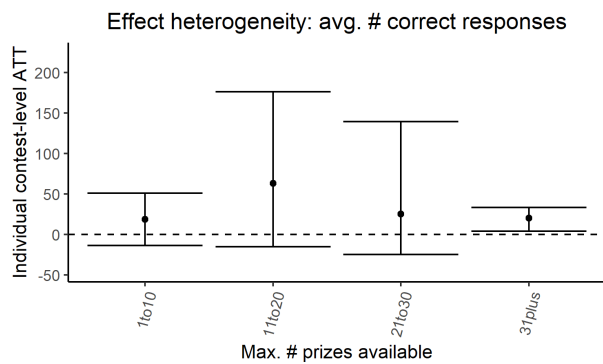


Figure 11: Conditional ATT based on number of prizes available

Figure Notes: Point represents Conditional ATT, and lower/upper bounds represent the 2.5th/97.5th percentiles of contest-level individual treatment effects

Finally, Figure 12 shows how treatment effects differ across contests by total prize basket size. Several effects of the streak-based evaluation metric are driven by contests with a high total basket value. Similarly to minimum prize value, the overall positive effect on percentage of responses qualified is reversed for high-basket contests.

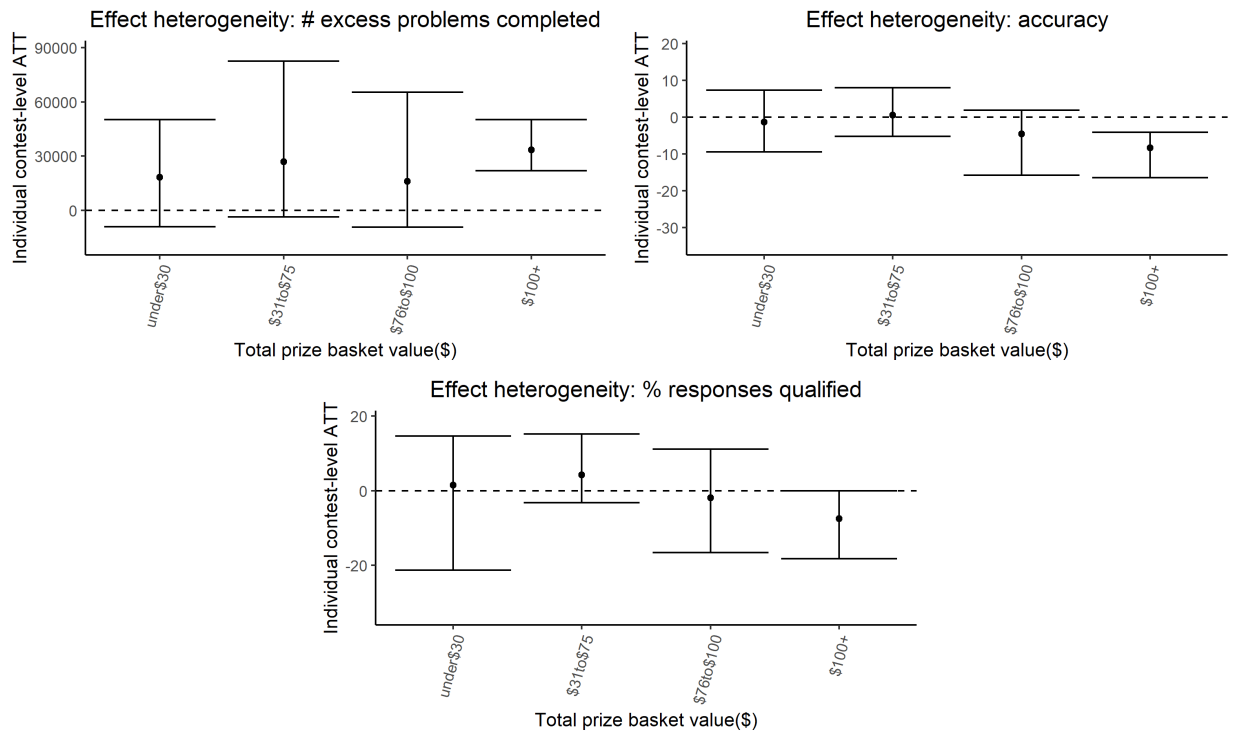


Figure 12: Conditional ATT based on total prize basket value

Figure Notes: Point represents Conditional ATT, and lower/upper bounds represent the 2.5th/97.5th percentiles of contest-level individual treatment effects

In summary, our main results about the effects of a streak-based evaluation metric on participant behavior are mainly present when the stakes are relatively high, *i.e.* when contests have a relatively high minimum prize, or total basket prize value, or when many prizes are available. In the case of participation, it seems streak contests might even be more successful at attracting participants if the minimum prizes are sufficiently high, increasing participation by 89.3%.

3.6 Discussion & Concluding Remarks

This paper studies the impact of using a streak-based evaluation metric on the performance of crowdsourcing contests. Overall, streak contests increase by 11.9% the *number* of responses and by 3.5% the *percentage* of responses in a contest that are qualified, *i.e.* the percentage of labels that are considered trustworthy. This is what matters most to the platform in the immediate term. It seems to be largely driven by the effectiveness of streak contests

in driving engagement beyond the minimum required level; participants in streak contests provide 83% more responses in streak contests than in accuracy contests, and 947% more responses above the minimum participation threshold.

For continuity purposes, crowdsourcing platforms also care about attracting large numbers of participants and attracting more skilled and high-performing participants. In addition, for some crowdsourcing platforms (including *CentaurLabs*), it is important to complete labels rapidly.

Table 15 provides a summary of our results:

Outcome variable	Control Mean	Impact of streak metric (+/-)*	Deg. significance	Heterogeneity?
Participation	208.6	None		Increase for high-minimum-prize contests
Normalized participation	3.3 %	None		
# problems completed	25,296	14% (+)	***	Increase most pronounced in high-minimum contests
Avg. # problems completed / participant	105.7	83% (+)	***	
# excess responses	13,577	133% (+)	***	Increase most pronounced in high-minimum and large-basket contests
Avg. # excess responses / participant	15.4	947% (+)	***	
Median response duration	4.1	-0.61% (-)	***	
Accuracy	61.32	-2% (-)	***	Decrease most pronounced in high-minimum and large-basket contests
Avg. # correct responses	49.2	102% (+)	***	Increase most pronounced in high-minimum contests with many prizes
% responses qualified	69.4	4% (+)	**	Increase reversed in high-minimum and large-basket contests
# responses qualified	13,844	12% (+)	***	Increase reversed in high-minimum and large-basket contests
Net labeling rate	10.8	31% (+)	**	N/A

Table 15: Summary of results.

Table Notes: *(+) indicates an improvement relative to the control mean, while (-) indicates decline. **Three stars indicate significance at $\alpha = 0.01$, two stars indicate significance at $\alpha = 0.05$, and one star indicates significance $\alpha = 0.1$

We find that streak contests do not necessarily outperform on several of these continuity metrics — for example they do not attract more participants — but they perform well

on others, (*e.g.* they induce participants to complete many more problems, and to do so faster). This increase in speed has several implications; in addition to allowing for greater efficiency (*i.e.*, the same amount of data can be labeled with fewer contests and thus at a lower cost), higher rates of responses and consensus may unlock new applications for collaborative labeling. For example, with sufficient speed, this and similar platforms could be used to provide second medical opinions (or other labels) directly to patients (or other end consumers) either in real-time or with a timing guarantee.

It is striking that streak contests are very successful at motivating participants to put in more effort once they have entered a contest, increasing both the volume and velocity of activity. The greater role of uncertainty under the streak evaluation metric means that participants must often complete many more problems in order to remain competitive, which they may partially compensate for by spending less time on each problem.

Individuals also complete more problems correctly in streak-based contests. While the percentage of correct problems decreases under the streak evaluation metric (partially driven by the significant increase in total problems completed), the additional incorrect problems are largely not incorporated into final consensus labels. The result is that contests based on a streak evaluation metric manage to produce net positive results on a platform-wide level, increasing the percentage of responses that fall above a quality threshold, and increasing the speed of consensus-based labeling. However, this suggests that the effectiveness of an evaluation metric that primarily drives engagement is contingent on the use of an effective aggregation algorithm.

The heterogeneous nature of the effect of the streak evaluation metric suggests that certain aspects of contest design, such as minimum prize value, interact significantly with these findings. Contests designers who decide to use a streak-based evaluation metric should take into account that such an evaluation metric is thus mainly effective when the minimum prize or total prize basket size are relatively high, or when many prizes are available.

Performance streaks have historically been seen as a statistical phenomenon that biases subsequent predictions or confidence levels (Gilovich et al. 1985), a “fun” tool to motivate participants (Huynh et al. 2018), or a type of “asset” that modifies participants’ risk-seeking behavior over time (Levy 1996); they have rarely been implemented as a basis for competitive

compensation. Our results suggest that in addition to serving as a useful psychological tool to keep participants returning to a platform, performance streaks—when tied to a platform’s compensation mechanism—can also improve the output generated by the platform.

There may be other secondary effects of streak-based evaluation metrics, for example on the ability of participants to learn and retain skills. Although such participant-level outcomes were out of scope for this analysis, they should be examined as an additional input into an overall strategy for using streak-based evaluation metrics.

The increased engagement, performance, and speed induced by streak evaluation metrics are important in the healthcare context studied by this paper, as they suggest that such metrics can increase the ability of crowdsourcing platforms to contribute to high-quality care by leveraging the knowledge of people around the world. More broadly, these results open up new options for structuring crowdsourcing incentives, which might be used by other distributed-task platforms in order to influence the behavior of participants.

Based on these results, designers of crowdsourcing tools may want to reconsider the value of performance streaks that are typically only used for gamification. Gamification is often used to influence participant activity via participant motivation; however, our results have shown that in addition to influencing participation and engagement metrics, gamification elements may also influence performance metrics when incorporated into the evaluation structure. The use of gamification features as performance drivers may both attract more diverse participant bases to crowdsourcing platforms and enable new applications of crowdsourcing in domains where participant incentives are not directly aligned with those of contest organizers.

3.7 Appendix

3.7.1 Explanatory variables

Category	Variable	Description
Format	Contest format	Scoring method e.g., <i>Streak</i> : winner has longest unbroken streak of correct answers <i>Accuracy</i> : winner has highest percentage of correct answers
	Content type	Text/ image/ video
	Classification complexity	Avg. number of answer choices per problem
Process	Contest duration	Hours between contest start and end (advertised in advance)
	Total prize basket	Sum of all available prizes in dollars (advertised in advance)
	Maximum prize value	Dollar value of prize for 1st place winner (advertised in advance)
	Minimum prize value	Dollar value of lowest prize granted (advertised in advance)
	# of winners	Number of prizes available (advertised in advance)
	Evenness/spread of prizes	Standard deviation of prize basket
Restrictions	Cap on # problems	Binary variable indicating the existence of a cap on number of permitted problem attempts(advertised in advance)
	Time limit	Maximum time (in seconds) allotted to answer each question
	Participation threshold	Minimum activity (number of completed problems) required to reach leaderboard (advertised in advance)
Environment	Contest order	Indicates how "early" or "late" contest occurred compared to other contests in the same topic group
	Contest purpose	Describes whether the contest was launched to label a client dataset, to keep participants engaged on the platform, etc.
	Audience	Indicates whether contest is targeted at a subset of participants (e.g., those who have never won, or cohort-based groups)
	Platform "crowdedness"	# of other contests available at start of contest
	Percentage "free" problems	Percentage of problems which are unlabeled at least once during contest
	Time trends	Month-year of contest start (to control for time-based unobservables)

Table 16: Available contest-level variables of interest

3.7.2 Alternate matching results

This section presents matching results and balance statistics for outcome measures which were only relevant to a subset of observations (see Table 4). For most outcomes, we excluded contests that attracted zero participants, yielding a dataset with 1477 observations (with one exception: for “avg. past ranking of participants” one additional observation is excluded, resulting in a sample size of 1476). For the $n = 1476$ and $n = 1477$ data subsets, we use the same matching protocol described in Section 3.4.2. Matching results and balance statistics for this subset are in Tables 17 and 18.

	# Contests		
	Streak (treatment)	Accuracy (control)	Total
All topic groups (103 topics)	319 (318)	1158	1477
Discarded due to lack of common support	0	445	445
Discarded due to exact, caliper, or kNN matching	89	378	467
# contests included in final sample	230 (229)	335	565

Table 17: Summary of matching process data sample for outcomes with $n = 1477$ (1476)

The outcome variable “net labeling rate” uses a sample size of $n = 1117$, consisting of those contests which did not overlap temporally with other contests in the same topic. When conducting matching for net labeling rate, we instituted a stricter set of calipers than was used for other outcomes. Specifically, in addition to the three covariates for which caliper restrictions were used when matching on the main dataset (total prize basket size, platform crowdedness, and number of winners), we required additional calipers for contest duration and contest order. Matched control contests must have a duration within one hour of the relevant treated contests, and must have a “contest order” value within one unit of the corresponding value for the relevant treated contests. These restrictions were necessary because 1) contest duration is a direct input into labeling rate, and it is thus particularly important to ensure closeness on this dimension, and 2) contest ordering may affect the number of problems available to be labeled. We used Mahalanobis distance rather than propensity scores to select matched pairs in order to ensure direct comparison on these

		<i>Before matching</i>			<i>After matching</i>			
	Variable	μ_T	μ_C	SMD _t	μ_T	μ_C	SMD _t ^m	
Format	Content type: image	0.76 (0.75)	0.68 0.68	0.17 0.17	0.72 0.72	0.72 0.72	-0.00 -0.00	
	Content type: text	0.12	0.09	0.09	0.12	0.12	0.00	
	Content type: video	0.13	0.23	-0.31	0.17	0.17	0.00	
	Answer choices per problem	3.75 (3.76)	4.04 4.04	-0.15 -0.15	3.68 3.68	3.69 3.69	-0.00 -0.00	
Process	Contest duration (hrs)	36.64 (36.68)	24.19	0.38	33.66 (33.70)	32.41 (32.03)	0.04 (0.05)	
	Minimum prize value: \$0.5	0.62 (0.63)	0.60	0.04	0.70 (0.71)	0.70 (0.71)	0.00 0.00	
	Minimum prize value: \$1	0.24 0.24	0.38 0.38	-0.32 (-0.33)	0.22 0.22	0.22 0.22	0.00 0.00	
	Minimum prize value: > \$1	0.13	0.01	0.35	0.07	0.07	0.00	
	Maximum prize value	14.60 (14.62)	17.06	-0.23	14.78 (14.80)	14.76 (14.72)	0.00 (0.01)	
	Evenness/spread of prizes	3.64 (3.65)	4.57	-0.28	3.96 (3.97)	4.04 (4.03)	-0.02	
	Maximum number of winners	17.48 (17.49)	20.74	-0.40	18.27 (18.29)	18.27 (18.28)	0.00 0.00	
	Total prize basket (\$)	48.83 (48.89)	65.35	-0.52	53.11 (53.21)	51.39 (51.30)	0.05 (0.06)	
	Restrictions	Cap on # problem attempts*	7687.8 (7680.50)	5766.34	0.46	7133.04 (7120.52)	7130.8 (7118.28)	0.00
		Per-question time limit (s)**	3602.7 (3595.14)	4839.5 (4839.51)	-0.42 -0.42	3454.8 (3443.65)	3409 (3397.64)	0.02 0.02
Environment	Contest order	13.23 (13.29)	17.27	-0.35 (-0.34)	15.48 (15.54)	14.68 (14.73)	0.07 0.07	
	Purpose: customer	0.39	0.76	-0.75	0.42	0.42	-0.00	
	Purpose: engagement	0.61	0.22	0.79	0.58	0.58	0.00	
	Purpose: research	0.00	0.02	-0.17	0.00	0.00	0.00	
	Audience: cohort	0.08	0.05	0.10	0.02	0.02	0.00	
	Audience: standard	0.92	0.95	-0.10	0.98	0.98	0.00	
	Platform "crowdedness"	6.73 (6.74)	6.77	-0.01	6.05 (6.07)	6.13 (6.14)	-0.02	
Percentage "free" problems	0.24	0.36	-0.49	0.26 (0.25)	0.27	-0.07		

Table 18: Pre/Post-matching balance summary for contest-level covariates (n = 1477(1476))

Table notes: μ_T and μ_C are the mean across treatment contests and control contests respectively.

*Contests with no cap are coded with a large value (10,000) for matching purposes only. **Contests with no per-question time limit are coded with a large value (6,000) for matching purposes only

covariate dimensions, and thus improved balance (although use of a propensity score does not change our results). Finally, we excluded “audience group” as a covariate in the matching and regression analyses due to lack of variation on this variable within the data subset. Matching results and balance statistics for this subset are in Tables 19 and 20.

	# Contests		
	Streak (treatment)	Accuracy (control)	Total
All topic groups (99 topics)	193	924	1117
Discarded due to lack of common support	0	477	445
Discarded due to exact, caliper, or kNN matching	105	795	900
# contests included in final sample	88	129	217

Table 19: Summary of matching process data sample for outcomes with $n = 1117$

	Variable	<i>Before matching</i>			<i>After matching</i>		
		μ_T	μ_C	SMD _t	μ_T	μ_C	SMD _t ^m
Format	Content type: image	0.85	0.68	0.46	0.77	0.77	0.00
	Content type: text	0.07	0.08	-0.04	0.05	0.05	0.00
	Content type: video	0.08	0.24	-0.57	0.18	0.18	0.00
	Answer choices per problem	3.86	4.04	-0.09	3.45	3.45	-0.01
Process	Contest duration (hrs)	36.70	22.05	0.50	30.72	31.62	-0.03
	Minimum prize value: \$0.5	0.66	0.60	0.14	0.74	0.74	0.00
	Minimum prize value: \$1	0.24	0.40	-0.36	0.25	0.25	0.00
	Minimum prize value: > \$1	0.09	0.01	0.30	0.01	0.01	0.00
	Maximum prize value	12.12	17.37	-0.74	13.11	13.10	0.00
	Evenness/spread of prizes	2.73	4.61	-0.97	3.32	3.37	-0.03
	Maximum number of winners	17.30	21.19	-0.51	18.76	19.12	-0.05
Total prize basket (\$)	42.08	66.45	-0.92	49.89	50.36	-0.02	
Restrictions	Cap on # problem attempts*	7930.57	5718.46	0.55	6459.09	6449.86	0.00
	Per-question time limit (s)**	3028.70	4998.0	-0.66	3217.90	3150.70	0.02
Environment	Contest order	10.76	17.66	-0.69	11.77	12.69	-0.09
	Purpose: customer	0.27	0.79	-1.15	0.38	0.38	-0.02
	Purpose: engagement	0.73	0.19	1.20	0.62	0.61	0.02
	Purpose: research	0.00	0.02	-0.18	0.00	0.00	-0.02
	Platform “crowdedness”	5.89	6.48	-0.20	5.65	5.96	-0.108
	Percentage “free” problems	0.22	0.37	-0.59	0.25	0.26	-0.04

Table 20: Pre/Post-matching balance summary for contest-level covariates ($n = 1177$)

Table notes: μ_T and μ_C are the mean across treatment contests and control contests respectively.

*Contests with no cap are coded with a large value (10,000) for matching purposes only. **Contests with no per-question time limit are coded with a large value (6,000) for matching purposes only

Chapter 4

Competitive uncertainty vs prize value uncertainty: the impact of prize endogeneity on participant behavior in crowdsourcing contests

Olumurejiwa A. Fatunde

Center for Transportation and Logistics, Massachusetts Institute of Technology

4.1 Abstract

In crowdsourcing and other types of contests, prize structure is often designed in a way that creates incentives for participants. This paper contributes to the literature on contest theory and decision-making under risk by exploring how a change in the source of prize basket uncertainty—from the *probability* of winning a prize to the *amount at stake*—affects participant choices. We use data on 5,418 crowdsourcing contests for medical diagnosis to evaluate the impact of running a *pool* contest (in which all participants who meet a performance threshold split a prize basket evenly) instead of a standard *rank-order* contest in which the number and size of prizes are determined exogenously and announced in advance.

Under pool contests, we observe a 50.2% increase in the *number* of qualified responses and a 29.9% increase in the *percentage* of qualified responses, representing a net positive for the platform. Pool contests increase activity levels, driving a 19.3% increase in number of problems that the average participant completed above the minimum participation threshold. We find that pool contests result in a 16.8% increase in accuracy for the average participant but a 7% decrease in accuracy for top performers, suggesting that participants modify their effort in order to meet performance thresholds. Finally, pool contests result in a 27.4% increase in the speed of producing high-quality responses, but do not significantly affect the overall rate of reaching consensus. Our analysis suggests that pool contests are an effective tool for incentivizing effort; however, the selection of the performance threshold is critical.

4.2 Introduction

The proliferation of contest-like activities in domains as diverse as innovation, software design, research procurement, sports contests, admissions, and even legal and political activity has led to the development of a significant body of research on the determinants of various aspects of contest success.

The goal of this paper is to explore the role of prize basket uncertainty in shaping participants' decisions with regard to participation and effort, as well as their ultimate performance. Specifically, we examine a particular type of contest structure—one in which prizes are evenly shared by participants who meet a performance threshold—to study the effect of shifting the locus of uncertainty from the probability of winning to the amount at stake.

Contest theory is a broad class of literature within economics and related fields in which researchers study the conditions under which agents exert (costly) effort in the hopes of obtaining a reward (Dechenaux et al. 2014, Segev 2020).

Three canonical types of contests are studied in the contest literature: Tullock-style "lottery" contests, in which a prize is granted to a single individual based on the relationship between the individual's effort and total collective effort; all-pay auctions, in which only the highest bidder wins, but all bidders must pay their bidding amount (and thus participation is both irreversibly and costly); and rank-order tournaments, in which one or more participants

receive prizes based on their competitive ranking on some measure of output (Dechenaux et al. 2014).

Rank-order contests are notable in that they decouple the basis of compensation from the inputs provided by participants. Participants are not paid directly for the volume of their effort; instead, rank-order tournaments involve a nonlinear transformation from the distribution of inputs to a discrete earning distribution, distinguishing them from more traditional input-based payment schemes such as piece rates (Lazear and Rosen 1981). Lazear and Rosen (1981) argue that although piece rates successfully shift risk towards workers in a way that incentivizes greater effort for greater pay, rank-based compensation schemes may be preferable in cases where inputs are difficult to monitor or participants deviate from risk-neutrality.

Beyond these three core contest types, there exist several less common generalizations. For example, Cason et al. (2010) studies contests in which prizes are allocated proportionally to the output quality of each participant. In this paper, we examine yet another generalization, which we refer to as "pool" contests, whose structure lies on the spectrum between rank-order contests and proportional-prize contests. In rank-order contests, prize values are exogenous (Cason et al. 2010). In pure proportional-prize contests, prize values are endogenously determined by the efforts of individuals. In pool contests, prize values are endogenously determined, but a given participant's prize value depends on both that participant's own efforts and on the number and efforts of other participants. Specifically, all participants in a pool contest split a prize basket evenly if they meet a specified quality threshold. Pool contests therefore fundamentally modify the risks faced by participants. Halac et al. (2017) and Deck and Kimbrough (2017) compare pool contests (which they refer to as "equal-sharing" contests) to winner-take-all contests under two information-disclosure conditions in the context of a two-period innovation contest with two winners, and find that pool contests dominate winner-take-all contests when information about competitor performance is not known to participants. However, we are unaware of any study of pool contests in an open crowdsourcing setting. Pool contests have analogues in managerial settings; for example, in the firm context, there may be situations in which employees split a bonus pool subject to meeting performance targets.

We explore the impact of pool contests on contest-level outcomes on a platform where both standard rank-order and pool contests are available, and examine the implications of changes to those outcomes for our understanding of decision-making under risk.

4.2.1 Past literature

One stream of the contest theory literature has explored the role of various contest design elements and other factors, such as the number and structure of prizes, evaluation schemes, and the number and heterogeneity of participants, on participant decisions such as entry and effort expended.

Numerous theoretical (Moldovanu and Sela 2001, Kalra and Shi 2001) and experimental (Freeman and Gelber 2010, Lim et al. 2009) papers have studied the conditions under which a single prize is preferable to multiple prizes. In general, multiple prizes have been shown to induce higher efforts when participants are risk averse or heterogeneous, or when the cost of effort function is convex (Dechenaux et al. 2014, Szymanski and Valletti 2005, Chen et al. 2011). Field experiments have provided evidence that higher numbers of contest participants lead to reduced individual effort (List et al. 2020, Casas-Arce and Martínez-Jerez 2009). Myerson and Wärneryd (2006) predicted that population uncertainty (i.e., settings in which the number of participants is a random variable) results in lower aggregate effort investment.

Researchers have also examined contests with endogenous prizes, and particularly contests in which the prize amounts change based on aggregate effort. Chung (1996) finds that when the (single) prize increases with aggregate effort, a wasteful amount of excess effort can result. In our setting, additional effort by rivals increases the probability that a given participant's expected earnings will be lower, thereby creating negative spillovers (Dechenaux et al. 2014).

Another subset of contest research examines the role of information disclosure in shaping participant behavior.

To our knowledge, most research in this area has focused on disclosure about the number, skill level, or past performance of participants in the contest. Drugov and Ryvkin (2019) show that the effect of information disclosure in rank-order tournaments depends on the curvature of participants' marginal cost of effort. Ludwig and Lünser (2012) examined the impact of

disclosing information about the performance of competitors on subsequent participant effort and performance, finding that participants adjust their effort by reducing (increasing) their effort in the second stage if they were previously leading (lagging).

The research closest to ours is that of Boosey et al. (2020), who use a lab experiment to explore the effect of disclosing the number of participants on effort exerted by contest entrants. They find that disclosure has a strong positive effect on effort investment for participants with a high outside option. In addition to focusing on a different treatment, our setting is different in two key ways: first, entry decisions in our setting can take place at any point during the contest (rather than taking place simultaneously). Secondly, our setting makes use of data from a field setting instead of a lab setting, though our study is observational rather than experimental. We hope that future research will expand on this work with the use of field experiments that can specifically test for potential explanations that we are unable to rule out using the available data.

In our setting, participants are uninformed about the size of the competitor pool before, during, and after contests. Holding this uncertainty constant, we modify whether the prize distribution is exogenous (and announced in advance). Our empirical exploration thus focuses on uncertainty surrounding the *prize amount*.

Our paper has clear overlap with the information-disclosure literature; while the key modification in our case is a change from exogenously- to endogenously-determined prize numbers and amounts, the policy of “always disclosing all available prize information” effectively means that participants receive advance prize information in the exogenous case but not in the endogenous case. Our treatment can thus be seen as either a change in prize structure or an information treatment.

A final stream of research within contest theory concerns itself with the role of behavioral and cognitive biases related to social preferences (e.g., envy, inequality aversion) and risk aversion in shaping the behavior of contest participants.

There is extensive literature within economics on decision-making under uncertainty. For example, several papers in the contest literature explore the relationship between risk tolerance of participants and their decisions to enter (and exert effort in) a competitive contest rather than accepting a deterministic quantity-based payment (Cason et al. 2010,

Eriksson et al. 2009). Experimental research has shown that risk-averse individuals are less likely to enter tournaments when a risk-free alternative is available (Eriksson et al. 2009, Dohmen and Falk 2011), and that inequality-averse individuals similarly avoid competitive environments (Bartling et al. 2009, Balafoutas et al. 2012).

Eisenkopf and Teyssier (2013) find that the elimination of envy (by removing competition among participants) reduces *average* effort provision, while eliminating loss aversion by removing the potential to earn below a set reference point reduces the *variance* of effort provision. Shupp et al. (2013) also finds a relationship between loss aversion and lower effort provision.

4.2.2 Research setting

Rank-order contests have occasionally been used on platforms for coordinating distributed workers or crowdsourcing inputs from a large group. In this paper, we use one such application to compare rank-order and pool contests.

We work with data provided by *CentaurLabs*, a healthcare company that operates *DiagnosticsUs*, a crowdsourcing platform for medical diagnosis. The company uses a mix of labeled and unlabeled data (consisting of media content such as images, text, video, or audio files) and solicits medical labels on each piece of data through game-like contests that run on the platform. The platform is open to all, though some contests are targeted at subsets of participants. Participants include those with no medical background, those currently in training, those planning to train in the future, and those currently practicing medicine or working in a medicine-adjacent profession.

The platform evaluates participant performance on labeled content by comparing responses to pre-existing labels, and then uses this performance to determine how heavily to weight participants' contributions on previously-unlabeled data. Opinions on unlabeled content are aggregated via a "collective intelligence" algorithm which combines human intelligence with computer intelligence by learning about participants' skill levels over time and using that information to decide which responses to use to form consensus labels for each problem. Performance on labeled data also determines the competitive rankings that are used to allocate prizes to participants in rank-order contests.

Given that the algorithm aims to label vast datasets using the contributions of participants who will each only see a small part of this data, the goals of the contests on the platform are to 1) improve the performance of all participants and 2) retain high-skilled participants on the platform for extended periods, rather than simply identify the best entry out of a crowd for a single contest. In this paper, we measure how pool contests impact the first goal by measuring the treatment effect on the number and percentage of responses meeting a quality threshold.

The paper proceeds as follows: In Section 4.3, we describe the contests included in this analysis, as well as the explanatory and outcome variables of interest. In Section 4.4, we describe our methods and examine the differences between rank-order contests and pool contests using an expected utility framework in order to make predictions about relative outcomes. In Section 4.5 we explore the results of our causal analysis, and in Section 4.6 we consider the possible behavioral explanations that could be responsible for the observed results.

4.3 Data and Key Definitions

Our analysis focuses on crowdsourcing contests run between September 2020 and January 2022. This dataset includes information about contest parameters and aggregate information about contest performance summarized from participant-level data. Each observation in the dataset corresponds to a single contest.

In this setting, a *contest* is a time-limited window during which a participant can complete problems on a specific topic. A *problem*, or question, here refers to a single medical image for which a participant is asked to provide a label. Each *topic*, or clinical area such as "Brain Hemorrhage" or "Pneumothorax", is associated with a pool of problems. Within a particular contest, problems from the relevant topic pool are stochastically assigned to participants with the aim of minimizing repeat exposure to questions and maximizing coverage across *participants*, or participants. As a result, the order in which problems appear is participant-specific. A completed problem, or response, refers to a single answer by a single participant on a particular problem. For some problems, the correct answer, or label,

is known in advance. Performance on these "labeled" problems is used to evaluate participant skill level and determine how much weight to give each participant when aggregating responses to unlabeled problems. The percentage of problems in a contest which are unlabeled ranges from 7.1% to 100%, with the average contest consisting of 34.8% pre-labeled problems.

Each crowdsourcing contest requires participant to complete a specific type of task. The two most common types of tasks on the platform are "classification" problems, which require a participant to select from between two and eight multiple-choice options, and "segmentation" problems, which require a participant to visually identify a specific part of the image and draw a box or shape around the affected area, thus indicating an opinion on a question of clinical significance. Shapes are drawn by connecting several points, or vertices, to create a polygon. In this paper, we focus on the subset of contests that are based on segmentation problems because other contests types have limited sample sizes or limited variability in the treatment of interest. A breakdown of the inclusion criteria is provided in Table 1.

The most successful participants in each contest earn cash prizes based on performance. For contests in this sample, the total prize basket value per contest varies between \$20 and \$200, with individual prizes ranging from \$0.25 to \$100. The number of winners for contests in this sample ranges from 1 to 37. Before they decide whether to participate in a contest, participants receive information about the start and end dates of the contest, scoring guidelines, the minimum participation threshold, the cap on number of allowed problem attempts, and any tiebreaker rules.

Participants are unaware of the identity of other participants before a contest begins; however, as the contest progresses, some limited competitor information may become available. Specifically, once participants begin to surpass the minimum participation threshold, a dynamic "leaderboard" displays the rankings of the top participants, as well their estimated earnings (e.g., the prize each leading participant would win if they maintained their current rank through the end of the contest). Participants can also view the profiles of people listed on the leaderboard to learn more about their backgrounds, if the relevant participant have shared this information.

In addition to contests, the platform provides access to practice problems that are not timed or rewarded. Participants may access a wide range of practice problems on the platform at any time. Before or during a particular contest, participants may also opt to complete practice problems that specifically prepare them for that contest. Links to these problems are provided along with the contest instructions.

We exclude practice problems from the dataset given our focus on outcomes in the contest setting.

	Number of Contests
# Contests starting and ending between 1 Sep 2020 and 31 Jan 2021	5,418
Excluded groups:	
# Contests discarded (.e.g, due to setup errors)	393
# Contests run for irrelevant purposes (e.g. testing)	258
# Contests set up for institutions or "squads" of participants*	131
# Contests restricted to a certain audience	1,244
# Contest formats other than segmentation	2,003
# Segmentation contests with scoring schemes other than accuracy	7
# Segmentation contests with multiple answers	47
# Target contests	1,335

*These contests were excluded due to concerns about network effects and violation of SUTVA assumptions

Table 1: Exclusion criteria for contests on the platform

4.3.1 Description of alternative prize structures

In this paper, we focus on two different contest structures which differ in how they select and reward winners. In standard rank-order contests, there is a specific dollar prize amount associated with each final ranking position.

We are specifically concerned with the class of imperfectly discriminatory rank-order contests that employs "noisy ranking", in which factors other than a participant's effort level contribute to the final ranking (Fu and Lu 2011). The problem-level score (which is an input into the final ranking) is a function of effort and a random component, similar to the model developed by Lazear and Rosen (1981). In our case, performance may also be influenced by additional systematic components, such as participant skill level. Neither contest organizers nor participants have perfect information about participant abilities, although it is reasonable to assume heterogeneity in skill levels given the diversity in both medical training

and platform experience among participants. Lazear and Rosen (1981) highlight that in such cases, it may be beneficial to provide participants with opportunities to learn about their skill levels before they make an investment decision. Indeed, the existence of practice problems on the platform may partially fill this role.

Platform users face a two-stage decision process: first, they review the advertised contest parameters and choose whether or not to enter the contest. Secondly, conditional on having entered the contest, participants must decide how much effort to exert in pursuit of a prize.

The problem-level score for problem i completed by participant j in contest k is given by

$$score_{ijk} = 100 \left(\frac{|A_{ijk} \cap A'_{ik}|}{|A'_{ik}|} \right)$$

where $|A_{ijk}|$ is the area of the shape A_{ijk} drawn by participant j on the image presented in problem i , and $|A'_{ik}|$ is the area of shape A'_{ik} , representing the official answer.

Participant j 's overall contest-level score is simply the average over all individual problem scores:

$$CS_{jk} = \frac{1}{N_{jk}} \sum_{i=1}^{N_{jk}} score_{ijk} \quad (4.1)$$

where N_{jk} is the total number of problems completed by participant j in contest k .

In rank-order contests the total prize basket size, the available number of prize slots, and the specific prize assigned to each final ranking position are determined by the organizers and announced in advance. The values of CS_{jk} for all participants are ranked in descending order, and the pre-announced prizes are allocated to a pre-announced maximum number of winners.

In pool contests, all participants for whom $CS_{jk} > QT_k$ receive an even share of the prize basket, where QT_k is a pre-announced, contest-specific performance threshold (specified by a minimum contest score, ranging from 80% to 95%). Because the final prize distribution is endogenously determined by participants' participation and effort decisions, the number of winners is not known—or even determined—until the contest is over. Therefore, only the total prize basket size is announced in advance.

If no participants meet the performance threshold in a given contest, no prizes are awarded, and the organizers receive the benefit of any effort invested for free. It is also possible for no participant to meet the minimum participation threshold—and thus for no prizes to be awarded—in a rank-order contest.

In rank-order contests, participants face two sources of uncertainty in evaluating their expected earnings: their own skill levels, which may influence final ranking, and *competitive uncertainty* related to the number, relative skill level, and effort investments of other participants. The contest organizers share in participants' ignorance about true skill levels, but they have full information about the size of the competitive field at any given time.

Because thresholds are set independently of participant performance, winning a prize in a "pool" contest is within each participant's control (under the assumption that every participant is able to meet the minimum quality threshold given enough attempts). Therefore, competitor skill level is not a source of uncertainty in pool contests. However, number and effort level of competitors remain as sources of competitive uncertainty. In addition, individual skill level affects the level of effort required to meet the prize threshold.

Figure 1 shows the typical score and prize amount associated with each ranking position under both schemes, as well as the typical scores and number of winners that result under each threshold level. Participant performance in pool contests appears to cluster around the stated performance threshold, with the highest thresholds producing accuracy levels comparable with those of the top performers in competitive rank-order contests.

These descriptive comparisons suggest that average scores for the top 28 participants are higher for non-pool contests. This effect, which is mostly driven by pool contests with lower performance thresholds, may be explained by 1) uncertainty surrounding the "cutoff" score required to win a prize in a rank-order contest, and 2) the ability to win a larger prize through better performance in a rank-order contest. In other words, rank-order contests are more competitive.

Figure 2 displays an example of the information that participants see when comparing different contest options. There is no visible difference between rank-order and pool contests on the landing page of the platform, where multiple contest are displayed next to each other.

The prize allocation rules for each contest are visible to participants on the contest's

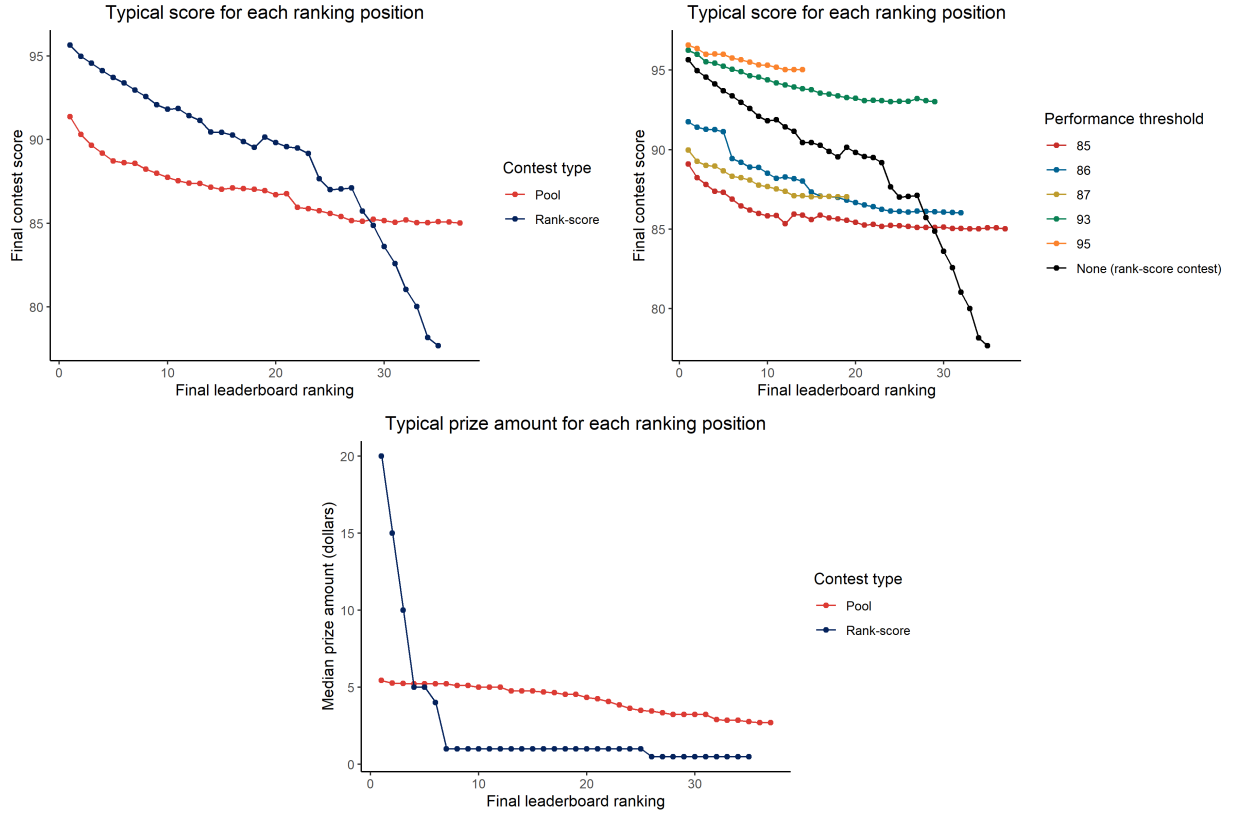


Figure 1: a) Representative scores for each prize structure, b) Representative scores for each prize structure with pool contests split by performance threshold, and c) Representative earnings for each prize structure

instruction screen (see Figure 3). On this page, participants can clearly see any minimum participation requirements, as well as the prize amounts that will be awarded to each ranking position (in a rank-order contest) or the criteria that participants must meet in order to receive a portion of the prize basket (in a pool contest).

For ranking position 5 or lower, the median prize amount is higher in a pool contest. This seems reasonable; for the most common prize basket sizes (\$80 and \$100), there would need to be 160 or 200 winners, respectively, to dilute the prize value in a pool contest to the same level as the most common prize amount for lower rankings (\$0.50) in a rank-order contest.

The number of winners in a pool contest generally decreases as the performance threshold increases, as expected.

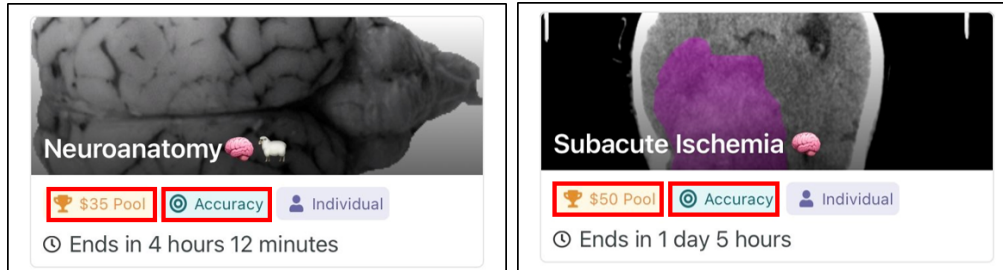


Figure 2: Example of information available to participants on the outer screen for a) rank-order and b) pool contests

4.3.2 Treatment variable

Our treatment variable indicates the presence or absence of uncertainty regarding N_k , the number of available prizes for a given contest k . Specifically, treatment is a binary variable which is equal to one for a “pool” contest in which N_k is endogenous and unknown before the contest starts, and equal to zero for a rank-order contest, in which N_k is exogenous and announced in advance.

4.3.3 Explanatory variables

For each contest, we have access to 1) “process” parameters such as the duration of the contest and the size of the prize basket, 2) restrictions or constraints (e.g., minimum number of completed problems required to reach the leaderboard or maximum permitted number of problem attempts, and 3) “environmental” variables which are not necessarily visible to (or salient for) participants, but should nevertheless be controlled for.

Table 2 defines the variables available in each of these four categories.

Table 3 provides descriptive statistics on each of these explanatory variables, and Figure 4 provides an indication of the level of activity on the platform over time.

Category	Variable	Description
Process	Contest duration	Hours between contest start and end (advertised in advance)
	Total prize basket	Sum of all available prizes in dollars (advertised in advance)
Restrictions	Cap on # problems	Binary variable indicating the existence of a cap on number of permitted problem attempts (advertised in advance)
	Time limit	Maximum time (in seconds) allotted to answer each question
	Participation threshold	Minimum activity (number of completed problems) required to reach leaderboard (advertised in advance)
Environment	Contest purpose	Describes whether the contest was launched to label a client dataset, to keep participants engaged on the platform, etc.
	Audience	Indicates whether contest is targeted at a subset of participants (e.g., those who have never won, or cohort-based groups)
	Platform "crowdedness"	# of other contests available at start of contest
	Percentage "free" problems	Percentage of problems which are unlabeled at least once during contest
	Time trends	Month of contest start (to control for time-based unobservables)

Table 2: Available contest-level variables of interest

Variable	N (%)	Mean (SD)
Contest format = "pool"	84 (6.3)	-
Contest duration (hrs)	-	20.03 (11.25)
Total prize basket (\$)	-	88.47 (28.71)
Contest has cap on # problem attempts	1060 (79.4)	-
Cap on number of problem attempts*	-	210.5 (89.73)
Problems have time limit	24 (1.8)	-
Per-question time limit* (seconds)	-	69.17 (35.41)
Participation threshold (problems)	-	73.67 (20.39)
Contest order within topic	-	71.33 (61.27)
Purpose: customer	1280 (95.9)	-
Purpose: engagement	55 (4.1)	-
Audience: standard	1151 (86.2)	-
Audience: cohort	184 (13.8)	-
Proportion "free" problems**	-	0.34 (0.18)
Platform "crowdedness"	-	7.56 (2.93)

*Average across contests for which restriction is relevant

**For contests with no participants/responses, values of these variables were imputed based on data from contests within the same topic

Table 3: Descriptive statistics for explanatory variables

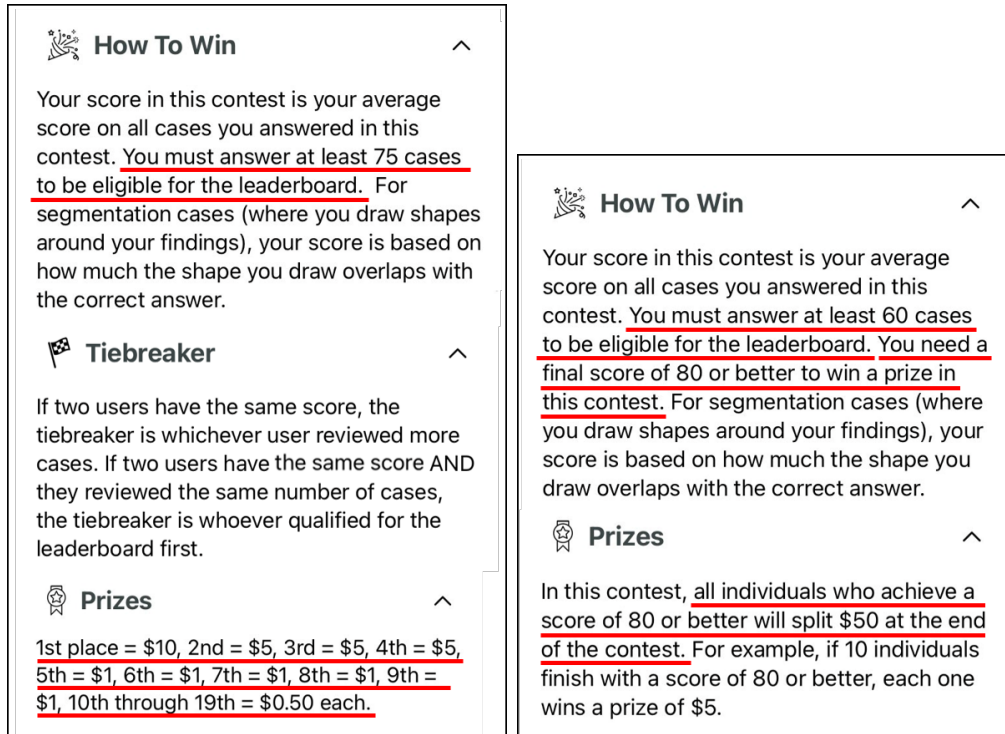


Figure 3: Example of information available to participants about the prize allocation within a) rank-order or b) pool contest

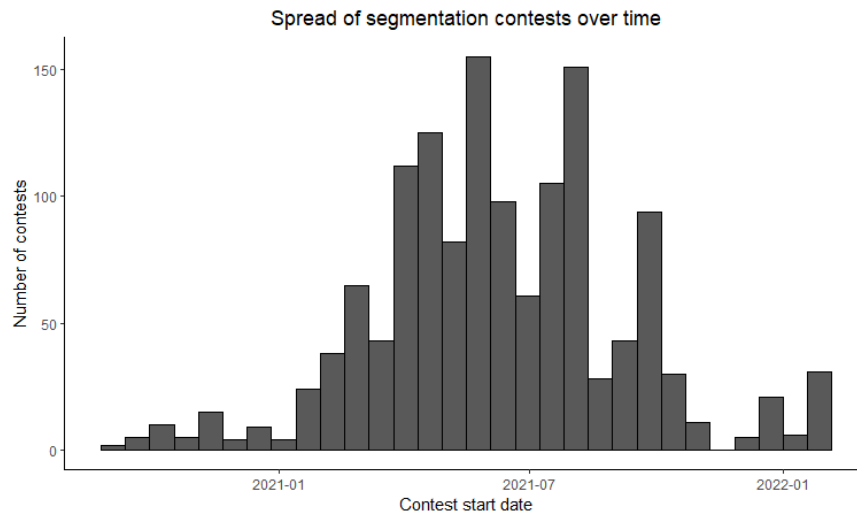


Figure 4: Level of activity over time

4.3.4 Outcome variables of interest

Given the goals of the collective intelligence algorithm, our outcome variables of interest fall into four categories: contest *participation*, which measures the number of participants,

conditional engagement, which measures the aggregate and per-problem effort exerted by participants after opting into a contest, *performance* metrics that measure a participant’s accuracy on problems for which a label exists, and speed, measured by time required to obtain responses/consensus labels. Table 4 describes each of the outcome variables in detail.

The primary outcomes of interest are the number and percentage of responses which exceed a pre-specified quality threshold. These metrics determine how many responses can serve as inputs to the final aggregated labels, and thus are central to the overall goal of the algorithm.

The remaining outcome measures influence these primary measures by changing either the number of qualified responses, the total number of responses, or both.

Some outcome measures use a subset of the full dataset. For example, outcomes that are conditional on participation exclude contests for which there were zero participants. Observations may also be removed if no participants met the performance threshold or in cases of missing data. For each outcome, the size of the pre-matching dataset is indicated by n in Table 4.

Category	Outcome	Description	n
Participation	Participation	# participants who completed ≥ 1 problem	1335
	Normalized participation	$100 \times \text{Participation} / \# \text{ active in past 15 days}$	1335
Engagement	# problems completed	# responses received from all participants	1291
	Avg. # problems completed/ part.	User-level problem count	1291
	# excess responses	Total # responses above min. threshold	1291
	Avg. # excess responses/ part.	# responses above min. participation threshold	1291
	Median response duration (sec)	Amount of time spent on a problem	1291
	Avg. # points drawn per image	# vertices in the polygon drawn by participant	1281
	Winning percentage	# prizewinners* per 100 participants	1291
Accuracy	Accuracy	Avg. contest-level score across labeled problems	1291
	Accuracy of top performers	Accuracy among participants who won prizes	1276
	# qualified responses	responses exceeding a quality threshold**	1291
	% responses qualified	# qualified responses / total # responses	1291
Speed	Qualified response** rate	# qualified responses** per hour	1291
	Net labeling rate***	$(\# \text{ new problems with consensus} - \# \text{ new problems without consensus})/\text{hr}$	881

*For rank-order contests, the number of prizewinners is not necessarily equal to the number of available prizes. If not enough participants meet the participation threshold, prizes may go unclaimed.

**The qualification threshold is a contest-specific minimum level of trailing accuracy

***Includes labels produced between the start time of the current contest and the submission time of the next contest’s first response

Table 4: Outcome variables of interest

4.4 Methodology

In order to understand the choice between rank-order and pool contests from the participant's perspective, we adopt an expected utility framework. In this case, net utility is represented by a participant's expected earnings minus effort expended.

$$\mathbb{E}[U_{jk}] = \mathbb{E}[\text{earnings}_{jk}] - c_j(\text{effort}_{jk})$$

where effort_{jk} is the effort expended by participant j conditional on entry into contest k , and c_j is a participant-specific constant that may include, among other things, participant j 's opportunity cost of taking part in the contest.

The calculation of expected earnings (and thus a participant's overall utility) is different for a standard rank-order contest than for a pool contest. We describe each of these in detail below.

4.4.1 Hypotheses

Expected earnings for a rank-order contest

If perfect information regarding a participant's skill level and the efforts of other contestants were available such that participants could precisely predict their expected final ranking, the expected earning for participant j in contest k would be determined by

$$\mathbb{E}[\text{earnings}_{jk}] = \hat{\ell}_{jk} P_{\ell,k} \tag{4.2}$$

where ℓ is participant j 's expected final ranking in contest k , and $P_{\ell,k}$ is the prize amount associated with that final ranking.

In most cases, however, participants are unable to precisely predict their ranking, partly because they have incomplete information about their own skill levels, and partly because final ranking is a function of the efforts of other participants. As such, participants may subconsciously estimate expected earnings using the following alternative construction:

$$\mathbb{E}[\text{earnings}_{jk}] = \sum_{\ell=1}^N p_{\ell,j} P_{\ell,k}$$

where N is the number of available prizes or rankings, $p_{\ell,j}$ is the participant's probability of attaining a certain final ranking, and $P_{\ell,k}$ is the prize associated with that final ranking.

This expression can also be written in terms of a particular contest-level score:

$$\mathbb{E}[\text{earnings}_{jk,rank-order}] = \sum_{\ell=1}^N P_{\ell,k} p(CS_{k,\ell} \leq X < CS_{k,\ell-1}) p(CS_{jk} \geq X) \quad (4.3)$$

where X is a random variable, $CS_{k,\ell}$ is the contest-level score corresponding to the ranking of ℓ in contest k , and $CS_{k,\ell-1}$ is the contest-level score corresponding to ranking $\ell - 1$. In other words, the second right-hand-side term captures the probability that score X falls between the cutoff scores for rankings ℓ and $\ell - 1$, and thus is sufficient to secure ranking ℓ (as determined by the competitiveness of the contest), and the third right-hand-side term captures the probability that participant j is able to attain a score of X (as determined by the participant's skill level). From the participant's perspective, there is uncertainty in both the second and third terms.

Because of the existence of the leaderboard, participants may form approximate (temporary) estimates of $p(\text{score}_{\ell} \leq X < \text{score}_{\ell-1})$ for various values of X during the contest by comparing these scores to the scores of the individuals on the leaderboard.

Expected earnings for a pool contest

For a pool contest, the participant's earnings depend not on his/her individual ranking, but on 1) the participant's own effort decision and resulting accuracy (which determine if the participant wins a prize) and 2) the collective effort exerted by other participants (which determines how big the participant's prize is).

Let

$$earnings_{jk} = \begin{cases} CS_{jk} \geq CS_{thresh,k} & \frac{T_k}{N_k p_{thresh,k}} \\ CS_{jk} < CS_{thresh,k} & 0 \end{cases}$$

where CS_{jk} is participant j 's overall score in contest k , $CS_{thresh,k}$ is the (pre-announced) minimum score required to enter the winner pool, T_k is the dollar value of the (pre-announced) prize basket, N_k is the number of contest entrants and $p_{thresh,k}$ is the proportion of entrants who met the performance threshold.

We can represent the expected earnings for a pooled contest as follows:

$$\mathbb{E}[earnings_{jk,pool}] = P_{pool,k} p(CS_{thresh,k} \leq CS_{jk}) \quad (4.4)$$

where

$$P_{pool,k} = \frac{T_k}{N_k p_{thresh,k}},$$

as described above. The second right-hand side term can be further broken down into two sub-terms in order to highlight the sources of uncertainty facing the participant:

$$\mathbb{E}[earnings_{jk,pool}] = P_{pool,k} p(CS_{thresh,k} \leq X) p(CS_{jk} \geq X) \quad (4.5)$$

where $p(CS_{thresh,k} \leq X)$ is the probability that a certain score X exceeds the performance threshold, and $p(CS_{jk} \geq X)$ is the probability that participant j is able to achieve a score of X .

In a pool contest, there is no uncertainty in the term $p(CS_{thresh,k} \leq X)$, since the score required to win a prize is a deterministic quantity which does not depend on the competitiveness of the contest. However, there is uncertainty in the prize amount $P_{pool,k}$, as well as the skill-level term $p(CS_{jk} \geq X)$.

We make the assumption that every participant is able to reach the performance threshold if he/she completes a sufficient number of problems. In practice, participants with higher skill levels will be able to achieve this score with a smaller number of problems, since they will get a higher percentage of problems correct earlier in their contest experience. Participants

with initially lower skill levels may need to spend time learning/practicing in order to reach a point where they are consistently getting high enough scores to translate to an average score of X . In addition, they may need to do additional problems to offset lower scores from the beginning of the contest. As such, lower-skilled participants will face a tradeoff between making a higher level of investment to secure an effectively “guaranteed” prize (of unknown value), taking their chances in a rank-order contest where they may be less likely to win a prize regardless of investment, or opting not to take part in a contest at all. On the other hand, higher-skilled participants may see pool contests as a less risky way to obtain a guaranteed payoff than a rank-order contest. The degree to which the tradeoff is worth it will depend on the size of the prizes available in the rank-order contest.

Because there is a theoretically unlimited number of competitors who *can* meet the quality threshold if they choose to make the required investment, the skill levels of competitors effectively disappears as a consideration for participants in pool contests. Instead, payoffs depend on own skill level (which determines the amount of investment required), as well as the collective effort put forth by other participants.

In summary, moving from one contest type to the other can be described as shifting uncertainty from the competitiveness of the contest (in a rank-order contest) to the size of the prize at stake (in a pool contest), as highlighted in Table 5.

	Rank-order contest	Pool contest
Prize amount		✓
Competitor skill level	✓	
Competitor effort level	✓	✓
Participant skill level	✓	✓

Table 5: Sources of uncertainty for participant j in contest k

Expected earnings under the two prize structures are equal when the following equality holds:

$$\sum_{\ell=1}^N P_{\ell} p(CS_{\ell} \leq X < CS_{\ell-1}) p(CS_{jk} \geq X) = \frac{T_k}{N_k p_{thresh,k}} p(CS_{jk} \geq CS_{thresh,k})$$

For a given participant, pool contests are thus made relatively more favorable by lower

thresholds, lower collective efforts of other participants, and a larger total prize basket. The relative attractiveness of rank-order contests depends on participant skill level; for a skilled participant, for whom $p(CS_{jk} \geq X)$ is larger for a given value of X , the attractiveness of rank-order contests increases with the value of the maximum prize and other high-ranked prizes. For a lower-skilled participant, rank-order contests become more attractive relative to pool contests when the value of the *minimum* prize increases. All of these comparisons may be distorted by participants whose decision-making is affected by cognitive biases such as risk aversion or ambiguity aversion.

Predictions

Sections 4.4.1 and 4.4.1 above, as well as exploration of previous literature, allow us to form hypotheses about the expected impact of pool contests on outcomes of interest.

Given the information above we predict that, relative to pool contests, rank-order contests with higher maximum prizes may remain an attractive option to high-skilled participants, while rank-order contests with lower maximum prizes will lose high-skilled participants to pool contests. We do not necessarily expect increased effort from these high-skilled participants, since the ability to win a guaranteed prize based on their current skill levels forms part of the appeal of pool contests for this group. The impact on participation decisions may differ for participants with different levels of risk aversion, although we do not directly measure risk aversion in this setting.

Given the factors that affect expected earning under each contest type, we expect that, conditional on entry, the change in effort for pool contest will depend on participant skill level. In a pool contest, participants need only exert enough effort to achieve the stipulated accuracy level, after which additional effort provides no additional boost to their expected earnings. This stands in contrast to rank-order contests, in which completing additional problems in a may give participants additional opportunities to improve their overall accuracy (and thus their rankings).

In aggregate, we expect to see increased effort in pool contests relative to rank-order contests, with the effort driven by lower-skilled or less experienced participants.

We expect the average accuracy in pool contests to be highly concentrated around

the performance threshold, since participants have no incentive to continue after meeting this threshold. Given the absence of competitive pressure to drive scores up, we expect that aggregate accuracy may decrease for pool contests with low performance thresholds. However, additional effort in pool contests may translate into a higher percentage of responses qualified.

We see no reason to expect a change in the speed of producing either qualified reads or consensus labels.

4.4.2 Causal effect of interest

Our goal is to measure the causal effect of shifting the locus of uncertainty from the probability of winning (i.e., a multi-winner rank-order contest) to the amount at stake (i.e., a pool contest).

The desired causal effect is a localized version of the Average Treatment effect on the Treated (ATT), which is given as

$$\hat{\tau} := \mathbb{E}[Y_k(1) - \hat{Y}_k(0) | T_k = 1].$$

where T_i is the treatment indicator, $Y_i(1)$ is the outcome under treatment and $\hat{Y}_i(0)$ is the estimated outcome under control. For treated contests, we observe the first term for each contest. The second term is not observed, and thus we use matching and simulation methods to identify counterfactual observations, impute these potential outcomes under control, and calculate a simulated version of the ATT for each outcome measure o :

$$ATT_{o,sim} = \frac{1}{|n_1|} \sum_{k=1}^{|n_1|} \left[[Y_{o,k} - \mathbb{E}[Y_{o,k}(T_k = 0)]] | T_k = 1 \right].$$

In this equation, n_1 refers to the set of treated contests and $|n_1|$ refers to the number of contests in this set. The corresponding values for control contests, n_0 and $|n_0|$, refer to the set of control contests and the number of control contests, respectively.

4.4.3 Matching

In order to ensure that treated and control contests are comparable, we use matching techniques.

We compare two matching approaches: a nearest-neighbor approach based on comparison of propensity scores, and the more flexible optimal full matching approach (Hansen and Klopfer 2006, Hansen 2004, Stuart and Green 2008). The optimal matching method returns a larger sample size; however, we proceed with the sample that results from nearest-neighbor matching due to slightly better balance results.

We require exact matching on topic group to ensure that only contests with a similar subject focus are compared. Within each topic group strata, we perform k-nearest neighbor matching with replacement, with “nearest” defined in terms of each contest’s propensity score. Propensity scores are calculated using logistic regression, with the indicator variable for pool contests (which is the treatment variable for our overall analysis) serving as the dependent variable. We include as covariates all explanatory variables which we expect to influence the probability of treatment. We exclude one variable, participation threshold, which is strongly tied to another variable (cap on permitted problem attempts), as well as number of current contests, since this is not considered when pool contests are launched. Other than the total basket size, we must exclude any summary measures related to the prize basket (e.g., minimum prize value, maximum prize value, prize spread), since these variables are not known in advance for pool contests, and thus represent post-treatment covariates.

We allow for variable-ratio matching, with each contest allowed between two and twelve control variables. Before matching, we discard control contests which fall outside the common support of the treated contests. Finally, we impose caliper restrictions on three variables in order to restrict matching to control contests which are sufficiently close on specific variables (and thus improve overall balance). We require matched control contests to be within \$5 of the total prize basket of corresponding treated contests, within four hours of the duration of treated contests, and within 30 percentage points of the percentage of free problems for treated contests. The caliper restriction for total prize basket is particularly important,

given that a key difference between rank-order and pool contests centers on the difference in participants' expected earnings.

The output of the matching procedure is a new data set that includes weights (w_k) for each on each observation. The weights on unmatched observations are equal to zero, which means that they will be eliminated from onwards analysis. w_k is equal to 1 for treated contests for which a match is found. For matched control contests, w_k values must satisfy two conditions; the weight of each individual matched control contest should be proportional to the number of treated contests to which it was matched, and the aggregate sum of control weights should equal the number of unique control contests that were used for at least one match:

$$w_{k|T_k=0} \propto \sum_{m \in n_1} \mathbf{1}(k \in M_m)$$

where M_m is the set of matched control contests for a given treated contest m , and

$$\sum_{k \in n_0} w_k = \sum_{k \in n_0} \mathbf{1} \left\{ \left[\sum_{k \in n_1} \mathbf{1}(m \in M_m) \right] > 0 \right\}.$$

In order to evaluate the success of matching, we compute the Standardized Mean Difference (SMD), a common measure of balance, both before and after matching. Before matching, the SMD is given by

$$\text{SMD}_t = \frac{\frac{1}{|n_1|} \sum_{k \in n_1} x_{k,t} - \frac{1}{|n_0|} \sum_{k \in n_0} x_{k,t}}{s_{\text{treated},t}}$$

After matching, the SMD incorporates the weights that are generated as part of the matching process:

$$\text{SMD}_t^m = \frac{\frac{1}{|n_1|} \sum_{k \in n_1} w_k x_{k,t} - \frac{1}{|n_0|} \sum_{k \in n_0} w_k x_{k,t}}{s_{\text{treated},t}}$$

where

n_0 and n_1 refer to the sets of control and treated contests, $|n_0|$ and $|n_1|$ represent the numbers of control and treated contests, and s_{treated} represents the sample standard deviation

for all treated units prior to matching (Ho et al. 2011).

We use the MatchIt package in R (Ho et al. 2011) to implement this procedure and select the counterfactual control contests that correspond to each treated contest. The matching process is summarized for the full dataset in Table 6; and for all sub-datasets in the Appendix. In this case, our treatment effect will differ slightly from the full ATT because a handful of treated observations were removed during the matching process. All subsequent mentions of the ATT refers to this reduced-sample treatment effect.

	# Topic Groups	# Contests		
		Treated (Pool)	Control(Rank-order)	Total
All contests with both streak and accuracy implementations	32	84	1251	1335
Discarded due to failure of common support		0	962	962
Unmatched implementations (due to exact, caliper, or kNN restrictions)		10	232	242
Matched implementations		74	57	131

Table 6: Summary of matching process data sample

Table 7 confirms that matching procedure was successful in creating balanced sets of treatment contests and counterfactuals. The SMD is at or below the threshold value of 0.1 for all covariates (Wang et al. 2013).

Variable	<i>Before matching</i>			<i>After matching</i>		
	μ_T	μ_C	SMD	μ_T	μ_C	SMD
Process Contest duration (hrs)	14.86	20.37	-0.73	12.97	12.94	0.00
Total prize basket (\$)	102.38	87.54	0.98	101.35	100.84	0.03
Restrictions Cap on # problem attempts*	305.26	2356.12	-1.91	191.11	196.62	-0.01
Per-question time limit (s)*	500,000	490,409	0.14	500,000	500,000	0.00
Environment Contest order	72.14	71.27	0.03	73.96	77.14	-0.10
Purpose: customer	0.98	0.96	0.12	1.00	1.00	0.00
Purpose: engagement	0.02	0.04	-0.12	0.00	0.00	0.00
Audience: cohort	0.00	0.15	-0.43	0.00	0.00	0.00
Audience: standard	1.00	0.85	0.43	1.00	1.00	0.00
Percentage "free" problems	0.26	0.34	-0.75	0.27	0.28	-0.05

*Contests with no cap are coded with a large value (10,000) for matching purposes only

*Contests with no per-question time limit are coded with a large value (500,000) for matching purposes only

Table 7: Pre/Post-matching balance summary for contest-level covariates

4.4.4 Effect estimation

After the matching process, we use regression methods to estimate treatment effects. Each regression uses the trimmed dataset that resulted from matching, as well as the corresponding matching weights. For each outcome variable, we use an appropriate regression method based on the outcome variable distribution. The selected outcome distribution for each outcome variable, chosen based on theoretical considerations and goodness-of-fit comparisons, is summarized in Table 8.

Category	Outcome variable	Outcome distribution
Participation	Participation	Negative binomial
	Normalized participation*	Gamma
	Avg. experience of participants (days)	Gamma
	Avg. past ranking of participants (%)	Gamma
	Avg. topic-specific experience (problems)	Gamma
Engagement	# problems completed	Negative binomial
	Avg. # problems completed per participant	Gamma
Engagement	# excess responses	Negative binomial
	Avg. # excess responses per participant	Normal
	Median response duration (sec)	Gamma
	Avg. # points drawn per image	Gamma
	Winning percentage*	Negative binomial
Performance	Accuracy	Normal
	Accuracy of top performers	Normal
	% responses qualified*	Normal
Speed	Qualified response rate	Gamma
	Net labeling rate	Gamma

*Transformed from percentage to rate over 100

Table 8: Statistical distributions used to model each outcome variable

Each regression takes the following general form:

$$Y_{o,k} \sim f_o(\theta_{o,k}, \sigma_o)$$

$$\theta_{o,k} = g_o(X_k, \beta_o)$$

where $Y_{o,k}$ represents the dependent variable for outcome variable o and $f_o(\theta_{o,k}, \sigma_o)$ represents the probability density function for the distribution corresponding to outcome o , with θ_o representing the systematic component and σ_o representing the ancillary parameter(s)

for the respective distribution (e.g., variance for Normal distribution, scale parameter for Gamma distribution, etc.). X_k refers to the $1 \times n$ covariate vector for contest k :

$$X_k = [1 \quad T_k \quad X_{1,k} \quad X_{2,k} \quad \dots \quad X_{(n-2),k}]$$

while β_o is a $n \times 1$ vector of corresponding coefficients.

For each regression, we are interested in the coefficient on the treatment indicator variable. For outcomes that are normally distributed, this coefficient is equivalent to the desired causal effect.

However, for non-normally distributed outcomes, the regression coefficient is uninformative with regard to the treatment effect. We use a simulation procedure developed by King et al. (2000) and Imai et al. (2008b) to gain an estimate of the treatment effect (and corresponding uncertainty measures) for these outcomes by comparing observed to imputed potential outcomes (Imbens 2004).

Using each set of regression results as an input, we take the parameter vector, $\hat{\delta} = [\hat{\beta}_o, \hat{\sigma}_o]$, and corresponding variance-covariance matrix, $\hat{V}(\hat{\delta})$, and use these to simulate a new set of parameters, $\tilde{\delta} = [\tilde{\beta}_o, \tilde{\sigma}_o]$, from the multivariate normal distribution. The simulated coefficients $\tilde{\beta}_o$ are combined with representative values of the contest-level covariates, X'_k , to compute new estimates of the systematic portion of the regression model. The values of X'_k are equivalent to those of X_k (observed values of covariates) for elements 1 and 3 through n . The second element of X'_k , which corresponds to the treatment indicator T_k , is set to 0 or 1, depending on whether we aim to generate expected values of the outcome variable under treatment or control. The simulated estimate of the systematic portion is given by:

$$\tilde{\theta}_{o,k} = g(X'_k, \tilde{\beta}_o)$$

Finally, a simulated estimate of the outcome variable for each regression model is computed by combining this simulated systematic portion with M draws from the distribution that represents the outcome variable. We choose $M = 1,000$ for this step. These repeated draws are necessary to account for fundamental uncertainty underlying the distribution of outcome measure o —that is, chance events that are not captured in X_k , but may nevertheless

affect the value of $Y_{o,k}$. For each $n = (1, \dots, M)$:

$$\tilde{Y}_{o,k}^n \sim f(\tilde{\theta}_{o,k}, \tilde{\sigma}_o).$$

We can obtain a single expected value of the outcome variable by taking the mean of all M simulated values of the outcome variable to average over fundamental uncertainty:

$$\tilde{\mathbb{E}}(Y_o) = \frac{\sum_{n=1}^{1000} \tilde{Y}_{o,k}^n}{1000}$$

This simulated expected value is designated as $\tilde{\mathbb{E}}(Y_o^{T_k=1})$ or $\tilde{\mathbb{E}}(Y_o^{T_k=0})$, depending on whether it is calculated with T_k counterfactually set to 1 or 0.

We repeat this simulation procedure 1,000 times, each time using a slightly different subset of the data generated via nonparametric bootstrapping. The 1,000 resulting estimates serve as the sampling distribution for the expected value of each outcome.

We use this procedure to simulate the difference between the observed value of treated observations and a simulated version of treated observations with the treatment variable counterfactually switched from 1 to 0 (Ho et al. 2011, Imai et al. 2012, Choirat et al. 2020, Imbens 2004), and with other covariates set to their observed values. The resulting simulated quantity gives us an estimate of the Average Treatment value for the Treated for our sample, defined as follows:

$$ATT_{o,sim} = \frac{1}{|n_1|} \sum_{k=1}^{|n_1|} \left[[Y_{o,k} - \tilde{\mathbb{E}}(Y_o^{T_k=0})] | T_k = 1 \right]$$

In Section 4.5, we present both the raw regression coefficients and the simulated value of the treatment effect for each outcome variable.

4.4.5 Heterogeneous Treatment Effects

In order to understand how treatment effects differ across certain contest-level characteristics, we compute heterogeneous treatment effects.

Given the nature of pool contests and our focus on earnings uncertainty, we are particularly interested in how treatment effects differ across the performance threshold for treated

contests, as well as the size of the prize basket.

We use one of a class of “meta-algorithms” proposed by Künzel et al. (2019) for flexibly estimating Conditional Average Treatment Effect (CATE)s. Specifically, we use the S-Learner, which is comparable to the “imputation estimator” described by Abadie and Imbens (2006).

In order to mirror the matched data sample as closely as possible, we use the trimmed dataset that resulted from the matching process.

We begin by estimating a response function using all observations from the post-matching dataset, $f_o(\theta_{o,k}, \sigma_o)$, using the distributions described in Table 8 and the same covariates as in the main regression. We estimate the response function jointly (rather than separately for treated and control observations) due to our limited sample size. The output is a set of estimated coefficients, denoted by $\hat{\beta}_o$. In order to impute the potential outcome under control for treated contests, we combine these coefficients with the observed covariates for treated contests, with the exception of T_k , which is counterfactually switched from 1 to 0. We therefore end up with

$$\hat{Y}_{o,k}^0 = g_o(X'_k, \hat{\beta}_o^{T_k=0}) \text{ for } k \in n_1$$

Using the imputation estimator, the estimated individual treatment effect for a particular treated contest is given by:

$$\hat{\tau}_{o,k} = Y_k - \hat{Y}_{o,k}^0 \text{ for } k \in n_1$$

The subgroup-level effect, or Conditional Average Treatment Effect on the Treated, for subgroup X (defined by a particular heterogeneity dimension) is therefore given by

$$\hat{\tau}_{o,X} = \frac{1}{|n_X|} \sum_{k=1}^{|n_X|} [\hat{\tau}_{o,k} | T_k = 1, k \in X]$$

where n_X is the number of contests in subgroup X . In Section 4.5, we show the CATT in the context of the distribution of individual contest-level individual treatment effects.

4.5 Results

Below, we show the simulated ATT representing the effect of a pool contest on each outcome measure.

4.5.1 Primary outcome metrics

Pool contests experience a significant and positive change in both the number and percentage of responses which surpass a quality threshold, as shown in Table 9 and Figure 5. As mentioned in earlier sections, these are the most important out of the outcome metrics, as they determine the success of the labeling process.

Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
% responses qualified	11.15	1.15	<0.001
height# responses qualified	0.45	0.05	<0.001

Table 9: Coefficients, standard errors and *p*-values for the treatment variable for primary outcomes

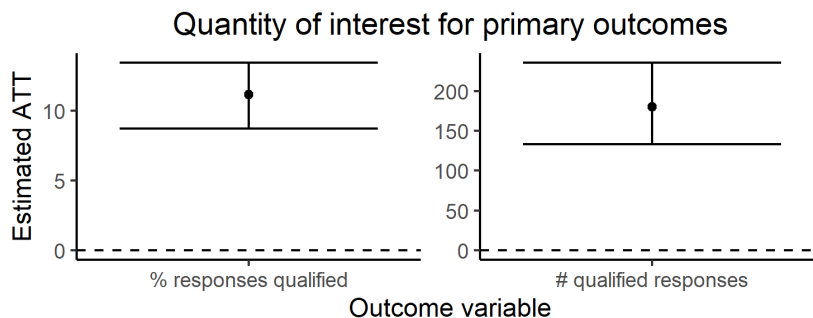


Figure 5: Simulated expected value of treatment effect for primary outcomes

4.5.2 Secondary outcome metrics

Participation

Pool contests attract a handful of additional participants, with the increase equal to $\approx 3-4\%$ of the participant population of an average rank-order contest; however, our results for

Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
Participation	0.05	0.02	0.04
Norm. participation	-0.02	0.03	0.48

Table 10: Raw regression coefficients for “participation” outcomes

normalized participation suggest that the participation increase may be explained by changes in the underlying participant base.

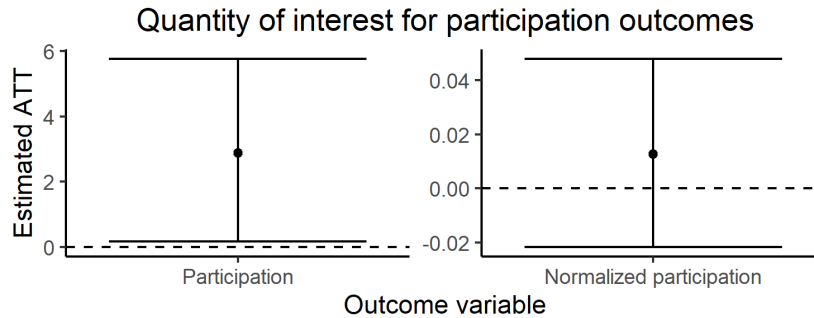


Figure 6: Simulated expected value of treatment effect

Some additional information about participant background may help us further understand the change in participation that occurs under pool contests. Out of 14,929 participants in the original dataset, 9,265 (62%) responded to a survey gathering information about their background and training. Figure 7 shows the average number of pool contests for each of the medical backgrounds reported in the survey.

The low average values in the top half of Figure 7 reflect the large number of participants who have not competed in pool contests. For the 936 survey respondents who completed at least one pool contest (shown in Figure 6b), the average number of pool contests completed per person ranges from less than three (for physician’s assistants and those with no medical experience) to greater than 6 (for medical students and medical doctors), and ranges from 20-33% of all contests completed.

Though we did not include medical training level as an outcome measure due to missing data for many participants, the relatively higher participation in pool contests of participants with some level of medical training suggests that these contests may draw participants with a higher level of overall familiarity (as represented by background/training).

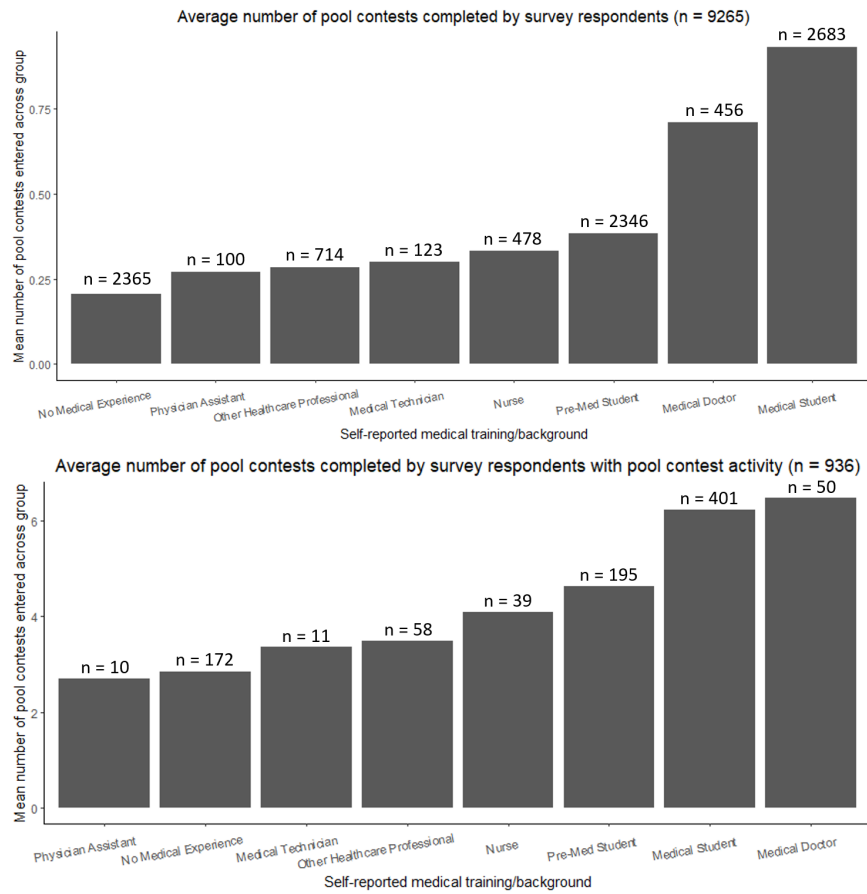


Figure 7: Average number of pool contests completed a) across all survey respondents and b) across survey respondents with at least one pool contest

Engagement

As expected, pool contests lead to an increase in engagement (conditional on participation), as measured by aggregate number of problems completed in each contest (655 additional problems, or a 27% increase relative to the mean control contest). This increase is driven by an average 8.43 additional responses by all participants (18.8 % increase), combined with a slight *decrease* in problems completed by participants who ultimately win prizes (this group completes 2.65 *fewer* problems in pool contests than in rank-order contests, representing a 3.3% decrease). This difference in conditional engagement of participants with different skill levels may reflect participants putting in only the level of effort required to meet the performance threshold; we expect this effort level to be higher for the average participant compared to top performers.

Notably, the mean number of excess responses is negative for both pool and rank-order contests in this sample, suggesting that a significant number of participants exit contests without exerting sufficient effort to win a prize. Pool contests make participants more likely to pass the minimum participation threshold, with participants completing an average of 4.27 additional excess responses—an increase of 19.3% relative to the mean control contest value of -22.1. This increase in engagement, together with the improvements in accuracy under pool contest (discussed in the next section), translates into an 18.5 percentage-point increase in the percentage of participants winning prizes in pool contests—an increase of 118% relative to the mean control contest. Pool contests result in prizes being distributed more widely, which is unsurprising given that the number of prizes awarded is endogenous.

Pool contests lead participants to draw an average of 1.79 fewer number of points drawn on each image (a 3.5% decrease relative to the mean control contest). However, there is no significant change in the amount of time spent on each response.

	Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
	# problems completed	0.24	0.04	<0.001
Avg. # problems completed per participant		-0.004	0.04	<0.001
	# excess responses	0.35	0.10	<0.001
Avg. # excess responses per participant		4.26	1.33	0.002
	Median response duration (sec)	-0.002	0.001	0.2
	Avg. # points drawn per image	0.002	0.0007	0.01
	Winning percentage	1.76	0.11	<0.001

Table 11: Raw regression coefficients for “engagement” outcomes

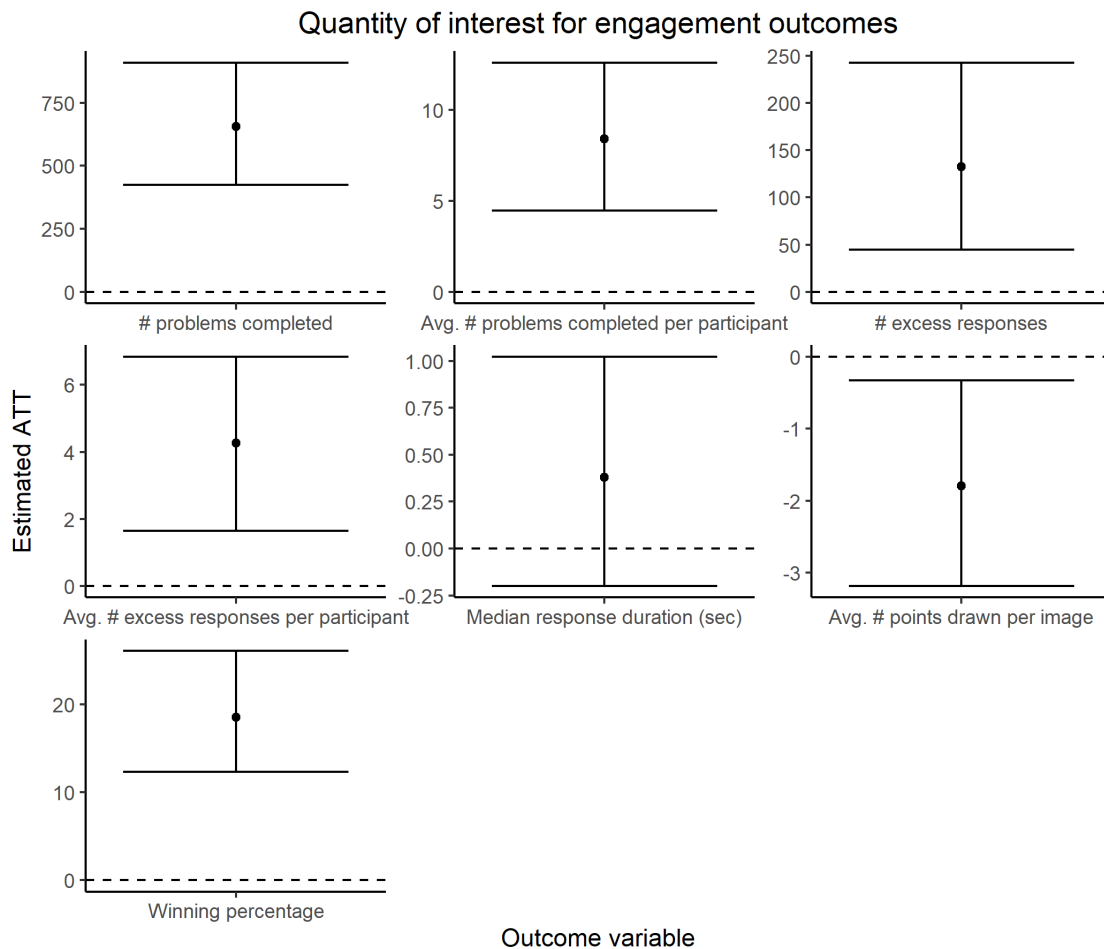


Figure 8: Simulated expected value of treatment effect

Performance

Pool contests have mixed effects on accuracy: they result in an increase of 7.5 points across all participants (an increase of 16.8% compared to the mean control contest). However,

accuracy of prizewinners is an average of 5.2 points lower (a 7% decrease) in pool contests. This is consistent with the descriptive trend shown in Figure 1, and is likely related to the less competitive nature of pool contests (as participants are not competing against other participants for limited prize spots), as well as the fact that the threshold for winning a prize is known in advance (which incentivizes participants to target this score and then stop after reaching it).

	Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
	Accuracy	7.52	0.80	<0.001
	Accuracy of top performers	-5.19	0.77	<0.001

Table 12: Raw regression coefficients for “performance” outcomes

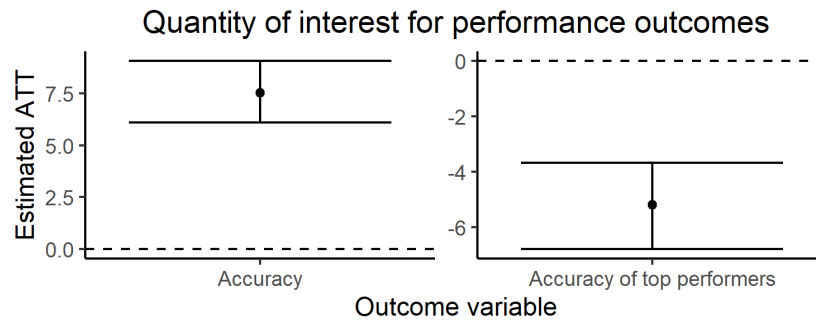


Figure 9: Simulated expected value of treatment effect

While the decrease in average scores for top performers is in line with our directional prediction for performance, the efforts of lower-ranked participants is sufficient to make up for this decrease, causing the overall increase in accuracy as well as an 11.1 percentage-point increase (an increase of 29.9% compared to the mean control contest) in the percentage of responses that pass a contest-specific quality threshold.

Speed

In order to evaluate how pool contests affect the velocity of contest activity, we measure 1) the production rate of qualified responses, and 2) the number of consensus labels produced per hour. We measure the labeling rate using a subset of the original data; specifically, we focus on contests which were not run simultaneously with any other contest drawing from the

same problem pool, since this allows us to isolate the contest’s contribution to the consensus labels for any given problem.

	Outcome	Coeff.	<i>s.e.</i>	<i>p</i>
	Net labeling rate	-0.07	0.04	0.07
	Qualified response rate	-0.02	0.005	0.001

Table 13: Raw regression coefficients for “speed” outcomes

Pool increase the speed of producing qualified responses by 3.29 responses per hour (a 27.4% increase relative to the mean control contest), but do not change the rate of problems achieving consensus status. This may be due to the fact that the performance thresholds drive some participants (particularly top performers) to target a lower level of accuracy than they would have attained in a rank-order contest.

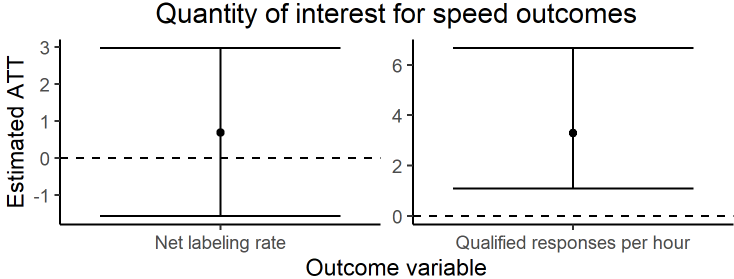


Figure 10: Simulated expected value of treatment effect

4.5.3 Effect heterogeneity

As discussed in section 4.4.5, we use the S-Learner procedure to determine whether the observed treatment effects vary based on the size of the prize basket or the level of the performance threshold. As summarized in Table 14, differences in performance threshold explain some of the differences in effect on performance outcomes, which is in line with our predictions. Treatment effects for outcomes from the participation, performance, and speed categories are heterogeneous across contests with different values of the prize basket size.

The (negative) effect on accuracy for top performers is driven by contests with the lowest performance threshold values (85%). This is in line with our predictions; as pictured in Figure 1, the median accuracy of top-ranked participants in rank-order contests is well

Outcome category	Effect of pool contest	Heterogeneity dimension	
		Performance threshold	Prize basket size
Participation	Attracts different mix of participants: newer, low-performers		✓
Engagement	Higher # completed problems, but fewer points drawn per image		
Performance	Increased accuracy and % qualified responses, but lower accuracy for top performers	✓	✓
Speed	More qualified responses, but same number of # labels per hour		✓

Table 14: Summary of effect heterogeneity.

Table Notes: checkmark denotes that the treatment effect varies along the corresponding heterogeneity dimension.

above 90% (and for first-prize winners, above 95%), meaning that these top performers could still expect to win prizes in pool contests with a low performance threshold even with a lower level of accuracy (and, presumably effort) than produced in comparable rank-order contests.

Contests with lower performance thresholds result in a higher percentage of participants winning prizes, as expected, while contests with higher thresholds see no change in the percentage of prizewinners.

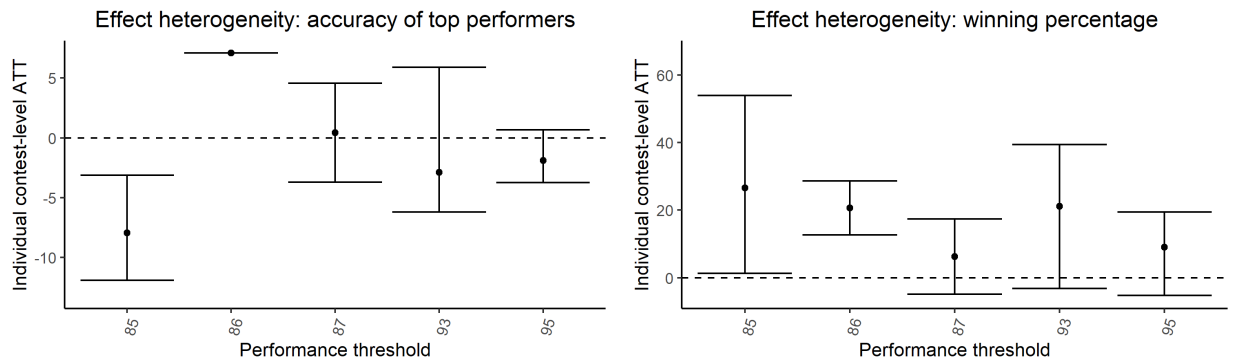


Figure 11: Conditional ATT based on minimum prize value

Figure Notes: Point represents Conditional ATT, and lower/upper bounds represent the 2.5th/97.5th percentiles of contest-level individual treatment effects

Several outcome measures exhibit heterogeneity based on prize basket size. Other than one contest with a basket size of \$60 (which we do not display here due to small sample size), the vast majority of contest have a basket size of \$100 or \$120.

The increase in accuracy and the increase in percentage of qualified responses are more pronounced in contests with larger baskets. This could be due to a combination of higher

thresholds and participants being motivated to exert greater effort given higher expected earnings.

The increase in the percentage of participants who earn prizes is reversed in contests with larger prize baskets, perhaps reflecting the positive relationship between prize basket value and performance threshold in pool contests.

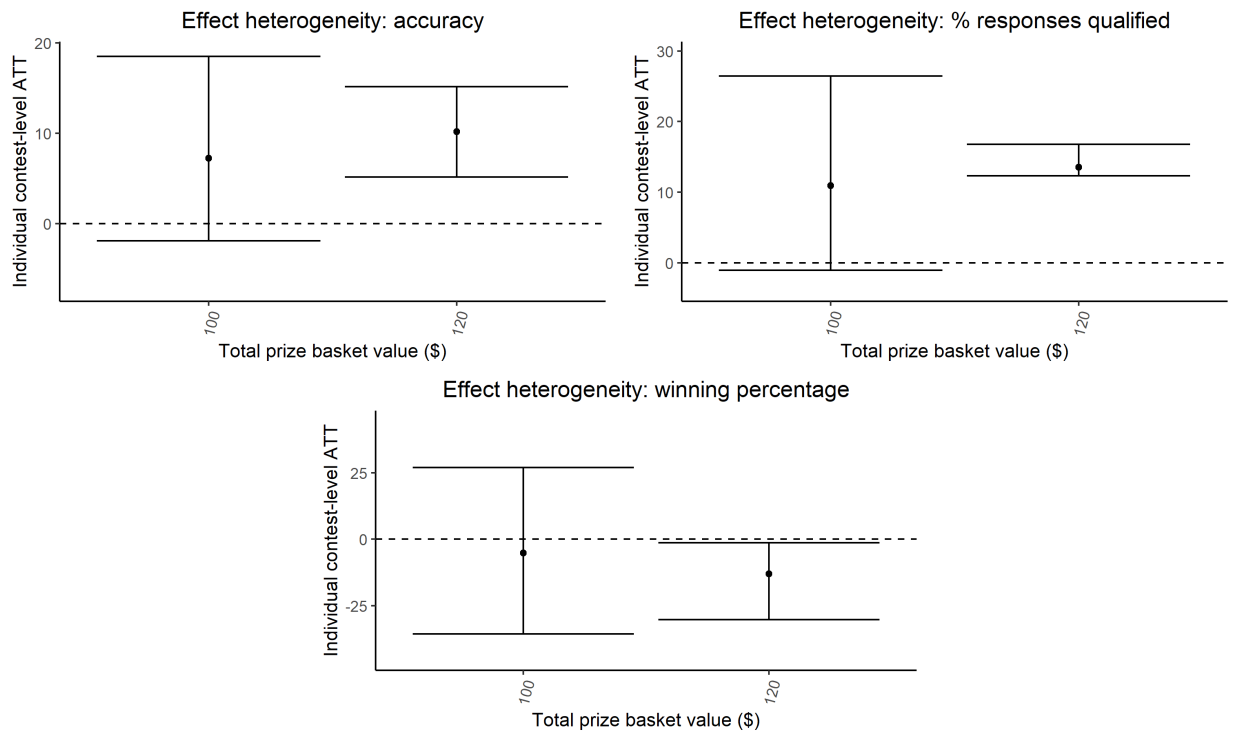


Figure 12: Conditional ATT based on minimum prize value

Figure Notes: Point represents Conditional ATT, and lower/upper bounds represent the 2.5th/97.5th percentiles of contest-level individual treatment effects

4.6 Discussion & Concluding Remarks

Our results are summarized in Table 15:

Overall, our analysis has demonstrated that pool contests incentivize participants to target their efforts in a manner that is consistent with meeting performance thresholds. The overall result is a positive for the platform: the number and percentage of responses exceeding a quality threshold both increase.

Pool contests also induce a greater number of participants to surpass the minimum

Outcome variable	Control Mean	Impact of streak metric (+/-)*	Deg. significance	Heterogeneity?
Participation	54.2	5.3% (+)	**	
Normalized participation	0.78 %	None		
# problems completed	2,404	27.2% (+)	***	
Avg. # problems completed / participant	44.95	18.8% (+)	***	
# excess responses	296.7	44.7% (+)	***	
Avg. # excess responses / participant	-22.1	19.3% (+)	***	
Median response duration	29	None		
Avg. # points drawn / image	51.41	3.5% (-)	***	
Winning percentage	15.7	117.8% (+)	***	Greatest improvement in low-threshold contests; effect reversed in large-basket contests
Accuracy	44.74	16.8% (+)	***	Greatest improvement in large-basket contests
Accuracy of top performers	73.95	7% (-)	***	Greatest decline in low-threshold contests
% responses qualified	37.1	29.9% (+)	***	Greatest improvement in large-basket contests
Qualified response rate	12	None ()	***	N/A
Net labeling rate	17.8	None ()	*	N/A

Table 15: Summary of results.

Table Notes: *(+) indicates an improvement relative to the control mean, while (-) indicates decline. **Three stars indicate significance at $\alpha = 0.01$, two stars indicate significance at $\alpha = 0.05$, and one star indicates significance $\alpha = 0.1$

participation threshold and increase accuracy for the average participant. The result is that a higher percentage of participants are rewarded with prizes in pool contests (particularly those with lower performance thresholds). For the vast majority of participants, these prizes are higher than what they would have earned in a rank-order contest. As a side effect, pool contests also cause a flattening of performance: by reducing the average performance of prizewinners and increasing the average performance of the general participant base, they reduce the differential in performance across the two groups.

From the organizers' perspective, the choice of performance threshold is an important parameter, since this threshold serves as a reference point for participants as they decide how much effort to expend. Performance thresholds are particularly important in guiding the behavior of skilled participants; contests with higher thresholds are able to produce accuracy levels comparable with those of the top performers in rank-order contests. The performance threshold (and other restrictive measures) thus play an important role in aligning the goals of individual participants with those of the contest organizers. Critically, if performance thresholds are set too low, they may "crowd out" the effort that participants may have exerted in a competitive rank-order contest.

From a behavioral perspective, there are multiple potential explanations for the observed changes in behavior under pool contests. One possibility is that participants are displaying inequality aversion, or a preference for "fair" contests in which all participants receive prizes of the same value. In nearly all cases in our sample, moving from a rank-order contest to a pool contest involves moving from unequal to equal prizes. Given the sample size limitations in our setting, we are unable to directly evaluate this theory by comparing participation and effort in rank-order contests with equal versus unequal prizes. We are therefore unable to disentangle any effect inequality aversion from participant calculations based on additional expected earnings. However, a fairness-based explanation would not be sufficient to fully explain the differences in effect for average versus top performers.

A second possibility is that the apparent change in information disclosure policy (i.e., a known versus an unknown number of winners) creates an ambiguity effect. However, we would expect participants with ambiguity aversion to shy away from pool contests. Instead, the targeted movement towards these contests suggests that these contests are considered to

be more favorable, despite the uncertainty surrounding payoff amounts.

A final possibility is that the change in behavior results from the simple fact of the prize structure changing from exogenous (in the rank-order contest) to endogenous (in the pool contest), which gives participants a greater measure of control over their eventual payoff and makes the relationship between engagement and expected earnings more salient.

This paper has provided empirical analysis of an alternative prize structure which is not frequently studied in the contest theory literature, but which has many potential applications. The ability to induce effort from participants by promising them an unknown share of a prize pool, particularly in the face of negative spillovers due to competitive effort, is powerful when combined with carefully selected performance thresholds. Pool contests also have equity implications, given their ability to spread prizes over a larger group without negative impact on overall quality or speed. Future experimental analysis of similar settings will help researchers to precisely identify the behavioral biases influencing individual decision-making in pool contests.

4.7 Appendix

4.7.1 Alternate matching results

The tables below present matching results for outcomes which are only relevant for a subset of the full dataset.

Many outcomes are only relevant for contests with positive participation, and thus have a sample size of $n = 1291$. One outcome (avg. past ranking of participants) excludes one additional observation, and thus has sample size $n = 1290$. Matching results and balance statistics for outcomes with sample size 1290 or 1291 are presented in Tables 16 and 17.

	# Topic Groups	# Contests		
		Treated (Pool)	Control(Rank-order)	Total
All contests with both streak and accuracy implementations	32	84	1207 (1206)	1291 (1290)
Discarded due to failure of common support		0	937 (936)	937 (936)
Unmatched implementations (due to exact, caliper, or kNN restrictions)		10	213	223
Matched implementations		74	57	131

Table 16: Summary of matching process data sample for outcomes with $n = 1291$ (1290)

One outcome, “accuracy for top performers”, is only relevant for contests in which at least one participant met the criteria for earning a prize. The analysis for this outcome is thus based on a dataset of size 1276. Matching results and balance statistics for this subsample are presented in Tables 18 and 19.

Finally, one outcome (net labeling rate) uses a subset of the data consisting of contests that do not overlap temporally with other contests in the same topic group, as these are the contests for which we can isolate their contribution to the labeling rate. For this data subset we modify the matching procedure slightly due to the smaller data size. We require exact matching on topic group, a discretized version of basket size, a discretized version of contest order, and a binary indicator of a cap on permitted problems. Within each strata we perform k-nearest neighbor matching with $k = 8$ and with nearest neighbors selected based on Mahalanobis distance. We also use calipers to restrict the distance between treated and matched control contests on contest duration (4hrs) and percentage of free problems (5.3

Variable	<i>Before matching</i>			<i>After matching</i>		
	μ_T	μ_C	SMD	μ_T	μ_C	SMD
Process Contest duration (hrs)	14.86	20.53 (20.48)	-0.75 (-0.74)	12.97	12.94	0.00
Total prize basket (\$)	102.38	87.63 (87.66)	0.98	101.35	100.84	0.03
Restrictions Cap on # problem attempts*	305.26	2321.25 (2314.88)	-1.88	191.11	196.62	-0.01
Per-question time limit (s)*	500,000	490,059 (490,051)	0.15	500,000	500,000	0.00
Environment Contest order	72.14	71.74 (71.8)	0.01	73.96	77.14	-0.10
Purpose: customer	0.98	0.96	0.13 (0.12)	1.00	1.00	0.00
Purpose: engagement	0.02	0.04	-0.13 (-0.12)	0.00	0.00	0.00
Audience: cohort	0.00	0.15	-0.43	0.00	0.00	0.00
Audience: standard	1.00	0.85	0.43	1.00	1.00	0.00
Percentage "free" problems	0.26	0.34	-0.73 (-0.74)	0.27	0.28	-0.05

*Contests with no cap are coded with a large value (10,000) for matching purposes only

*Contests with no per-question time limit are coded with a large value (500,000) for matching purposes only

Table 17: Pre/Post-matching balance summary for contest-level covariates (n = 1291 (1290))

	# Topic Groups	# Contests		Total
		Treated (Pool)	Control(Rank-order)	
All contests with both streak and accuracy implementations	32	75	1201	1276
Discarded due to failure of common support		0	931	931
Unmatched implementations (due to exact, caliper, or kNN restrictions)		7	214	221
Matched implementations		68	56	124

Table 18: Summary of matching process data sample for outcomes with n = 1276

Variable	<i>Before matching</i>			<i>After matching</i>		
	μ_T	μ_C	SMD	μ_T	μ_C	SMD
Process Contest duration (hrs)	14.93	20.51	-0.74	13.12	13.12	0.00
Total prize basket (\$)	102.67	87.78	0.93	101.47	100.85	0.04
Restrictions Cap on # problem attempts*	319.93	2323.61	-1.77	192.57	194.85	-0.00
Per-question time limit (s)*	500,000	490,010	0.15	500,000	500,000	0.00
Environment Contest order	72.81	72.07	0.02	74.75	77.17	-0.08
Purpose: customer	0.97	0.96	0.10	1.00	1.00	0.00
Purpose: engagement	0.03	0.04	-0.10	0.00	0.00	0.00
Audience: cohort	0.00	0.15	-0.43	0.00	0.00	0.00
Audience: standard	1.00	0.85	0.43	1.00	1.00	0.00
Percentage "free" problems	0.29	0.34	-0.71	0.29	0.29	0.04

*Contests with no cap are coded with a large value (10,000) for matching purposes only

*Contests with no per-question time limit are coded with a large value (500,000) for matching purposes only

Table 19: Pre/Post-matching balance summary for contest-level covariates (n = 1276)

percentage points). Matching results and balance statistics for this subset, which consists of 881 contests, are in Tables 20 and 21.

	# Topic Groups	# Contests		
		Treated (Pool)	Control(Rank-order)	Total
All contests with both streak and accuracy implementations	32	60	821	881
Unmatched implementations (due to exact, caliper, or kNN restrictions)		13	763	776
Matched implementations		47	58	105

Table 20: Summary of matching process data sample for outcomes with $n = 881$

Variable	<i>Before matching</i>			<i>After matching</i>		
	μ_T	μ_C	SMD	μ_T	μ_C	SMD
Process						
Contest duration (hrs)	14.57	19.22	-0.74	14.00	13.87	0.02
Total prize basket: under \$30	0.00	0.02	-0.14	0.00	0.00	0.00
Total prize basket: \$31 to \$75	0.02	0.17	-1.21	0.00	0.00	0.00
Total prize basket: \$76 to \$100	0.82	0.56	0.66	0.83	0.83	0.00
Total prize basket: \$101+	0.17	0.25	-0.22	0.17	0.17	0.00
Restrictions						
Cap on # problem attempts*	355	2030.88	-1.32	191.49	186.17	0.00
Per-question time limit (s)*	500,000	486,604	0.17	500,000	500,000	-0.00
Environment						
Contest order decile	5.32	5.77	-0.25	5.32	5.32	0.00
Purpose: customer	0.98	0.95	0.29	1.00	1.00	0.00
Purpose: engagement	0.02	0.05	-0.29	0.00	0.00	0.00
Audience: cohort	0.00	0.04	-0.22	0.00	0.00	0.00
Audience: standard	1.00	0.96	0.22	1.00	1.00	0.00
Percentage "free" problems	0.26	0.33	-0.71	0.28	0.27	0.10

*Contests with no cap are coded with a large value (10,000) for matching purposes only

*Contests with no per-question time limit are coded with a large value (500,000) for matching purposes only

Table 21: Pre/Post-matching balance summary for contest-level covariates ($n = 881$)

Chapter 5

Conclusion

5.1 Managerial Insights

The results presented in the preceding chapters provide insight into a small but important subset of the rapidly-expanding class of “informal supply chains” which operate without the constraints that govern traditional supply chains for products and services.

In the absence of formal ties, interpersonal relationships—and the implicit norms driving them—can often serve as the foundation of connections between supply chain actors. Chapter 2 shows that in the retail industry, these relationships are in fact pivotal to the performance of more vulnerable supply chain partners.

Contracts have often been used to send signals about private information (such as skill level or value) and coordinate the behavior of supply chain actors. In their absence, other tools are needed to induce desired behavior from upstream or downstream supply chain partners. Chapters 3 and 4 demonstrate that building incentives into contest compensation mechanisms can serve as an effective method of coordinating the participation and effort investment decisions of a large (potentially infinite) number of suppliers with heterogeneous backgrounds.

Informal supply chains often involve individuals or organizations that might have been excluded from participating in more formal supply chain arrangements (*e.g.*, due to lack of proper credentials or documentation). As such, this dissertation and other research that

explores potential levers for improved performance of informal supply chains may inform the many newly emerging and continuously evolving methods for organizing work and workers, particularly with the use of digital platforms.

5.2 Future Research Directions

In the retail sector, informality is likely a long-term, if not permanent, feature that distributors and manufacturers will need to navigate if they wish to reach customers at the Base of the Pyramid or other hard-to-reach populations. While our research in Chapter 2 focused on relationships with the retailers closest to these consumers, future research would benefit from collecting data directly from consumers to understand the factors driving their purchase decisions. Future research may also build on our findings about the disproportionate importance of relationships for informal retailers by testing various strategies for mitigating unavoidable disruptions to close business relationships.

Our work in Chapters 3 and 4 relied on assumptions about user motivations for participating in crowdsourcing contests. In future research, we plan to explore patterns of platform behavior to infer whether participants' "revealed motivations" are consistent with these assumptions (and with participants' stated motivations).

Chapters 3 and 4 took a cross-sectional approach to evaluating performance at a platform level, which did not allow us to draw conclusions about how skill levels and performance change over time (*i.e.*, learning). In future research we hope to be able to produce more insights by analyzing behavior at the user level.

From a methodological standpoint, this dissertation consists of observational studies based on historical data provided by research partners. While we were able to draw meaningful insights using quasi-experimental empirical methods, future research may be able to more precisely identify the behavioral mechanisms underlying participant decisions through the use of experimental methods.

In summary, many new directions remain to be explored in the realm of informal supply chains. We hope, through future research, to continue to develop actionable insights about these nontraditional, but potentially powerful, channels for delivery of products and services.

Bibliography

- Abadie A, Imbens GW (2006) Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–267, ISSN 00129682, 14680262, URL <http://www.jstor.org/stable/3598929>.
- Acimovic J, Parker C, F Drake D, Balasubramanian K (2020) Show or Tell? Improving Inventory Support for Agent-Based Businesses at the Base of the Pyramid. *Manufacturing & Service Operations Management* msom.2020.0922, ISSN 1523-4614, 1526-5498, URL <http://dx.doi.org/10.1287/msom.2020.0922>.
- Aklin M, Urpelainen J (2020) Trials and tribulations: Lost energy access gains in rural india. *Energy for Sustainable Development* 55:190–200, URL <https://doi.org/10.1016/j.esd.2020.01.002>.
- Anderson SJ, Kundu A, Ramdas K (2018) Disruptions, Resilience and Performance of Emerging Market Entrepreneurs: Evidence from Uganda. *SSRN Electronic Journal* ISSN 1556-5068, URL <http://dx.doi.org/10.2139/ssrn.3197806>.
- Andrikogiannopoulou A, Papakonstantinou F (2018) Individual reaction to past performance sequences: Evidence from a real marketplace. *Management Science* 64(4):1957–1973, URL <http://dx.doi.org/10.1287/mnsc.2016.2636>.
- Austin PC (2014) A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine* 33(6):1057–1069, ISSN 0277-6715, 1097-0258, URL <http://dx.doi.org/10.1002/sim.6004>.
- Austin PC, Stuart EA (2017a) Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical Methods in Medical Research* 26(6):2505–2525, ISSN 0962-2802, 1477-0334, URL <http://dx.doi.org/10.1177/0962280215601134>.
- Austin PC, Stuart EA (2017b) The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical Methods in Medical Research* 26(4):1654–1670, ISSN 0962-2802, 1477-0334, URL <http://dx.doi.org/10.1177/0962280215584401>.
- Balafoutas L, Kerschbamer R, Sutter M (2012) Distributional preferences and competitive behavior. *Journal of Economic Behavior & Organization* 83(1):125–135, URL <http://dx.doi.org/10.1016/j.jebo.2011.06.018>.
- Baquié S, Urpelainen J (2017) Access to modern fuels and satisfaction with cooking arrangements: Survey evidence from rural india. *Energy for Sustainable Development* 38:34–47, URL <http://dx.doi.org/10.1016/j.esd.2017.02.003>.
- Bartling B, Fehr E, Maréchal MA, Schunk D (2009) Egalitarianism and competitiveness. *American Economic Review* 99(2):93–98, URL <http://dx.doi.org/10.1257/aer.99.2.93>.

- Bigham JP, Bernstein MS, Adar E (2014) Human-computer interaction and collective intelligence URL <http://dx.doi.org/10.1184/R1/6470123.V1>.
- Boosey L, Brookins P, Ryvkin D (2020) Information disclosure in contests with endogenous entry: An experiment. *Management Science* 66(11):5128–5150, URL <http://dx.doi.org/10.1287/mnsc.2019.3488>.
- Boulaksil Y, Belkora MJ (2017) Distribution Strategies Toward Nanostores in Emerging Markets: The Valencia Case. *Interfaces* 47(6):505–517, ISSN 0092-2102, 1526-551X, URL <http://dx.doi.org/10.1287/inte.2017.0914>.
- Boulaksil Y, Fransoo JC, Blanco EE, Kouvida S (2019) Understanding the fragmented demand for transportation – Small traditional retailers in emerging markets. *Transportation Research Part A: Policy and Practice* 130:65–81, ISSN 09658564, URL <http://dx.doi.org/10.1016/j.tra.2019.09.003>.
- Calmon AP, Jue-Rajasingh D, Romero G, Stenson J (2017) Consumer Education and Regret Returns in a Social Enterprise. *SSRN Electronic Journal* ISSN 1556-5068, URL <http://dx.doi.org/10.2139/ssrn.2882402>.
- Casas-Arce P, Martínez-Jerez FA (2009) Relative performance compensation, contests, and dynamic incentives. *Manag. Sci.* 55:1306–1320.
- Cason TN, Masters WA, Sheremeta RM (2010) Entry into winner-take-all and proportional-prize contests: An experimental study. *Journal of Public Economics* 94(9-10):604–611, URL <http://dx.doi.org/10.1016/j.jpubeco.2010.05.006>.
- Chen DL, Moskowitz TJ, Shue K (2016) Decision making under the gambler’s fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics* 131(3):1181–1242, URL <http://dx.doi.org/10.1093/qje/qjw017>.
- Chen H, Ham SH, Lim N (2011) Designing multiperson tournaments with asymmetric contestants: An experimental study. *Management Science* 57(5):864–883, URL <http://dx.doi.org/10.1287/mnsc.1110.1325>.
- Chen PY, Pavlou PA, Yang Y (2014) Determinants of open contest participation in online labor markets. *SSRN Electronic Journal* URL <http://dx.doi.org/10.2139/ssrn.2510114>.
- Choirat C, Honaker J, Imai K, King G, Lau O (2020) *Zelig: Everyone’s Statistical Software*. URL <https://zeligproject.org/>, version 5.1.7.
- Chung TY (1996) Rent-seeking contest when the prize increases with aggregate efforts. *Public Choice* 87(1-2):55–66, URL <http://dx.doi.org/10.1007/bf00151729>.
- ClearTax (2021a) GST Registration Threshold Limits Increased. URL <https://cleartax.in/s/gst-registration-limits-increased>.
- ClearTax (2021b) Taxable Person Under GST. URL <https://cleartax.in/s/taxable-person-gst>.
- Collective GD (2019) Last Mile Distribution: State of the sector report. Technical report, Rugby, UK.
- Criscuolo P, Dahlander L, Grohsjean T, Salter A (2021) The sequence effect in panel decisions: Evidence from the evaluation of research and development projects. *Organization Science* 32(4):987–1008, URL <http://dx.doi.org/10.1287/orsc.2020.1413>.
- D-Lab M (2021) Prior-Year CITE Projects & Reports | MIT D-Lab. URL <https://d-lab.mit.edu/research/comprehensive-initiative-technology-evaluation/prior-year-cite-projects-reports>.

- Dechenaux E, Kovenock D, Sheremeta RM (2014) A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics* 18(4):609–669, URL <http://dx.doi.org/10.1007/s10683-014-9421-0>.
- Deck C, Kimbrough EO (2017) Experimenting with contests for experimentation. *Southern Economic Journal* 84(2):391–406, URL <http://dx.doi.org/10.1002/soej.12185>.
- Delechat C, Medina L (2020) What is the informal economy? having fewer workers outside the formal economy can support sustainable development. *IMF Finance Development* 54–55, URL <https://www.imf.org/external/pubs/ft/fandd/2020/12/what-is-the-informal-economy-basics.htm>.
- Deloitte M (2017) Reaching deep in low-income markets: Enterprises achieving impact, sustainability, and scale at the base of the pyramid. Technical report.
- Dohmen T, Falk A (2011) Performance pay and multidimensional sorting: Productivity, preferences, and gender. *American Economic Review* 101(2):556–590, URL <http://dx.doi.org/10.1257/aer.101.2.556>.
- Drugov M, Ryvkin D (2019) The shape of luck and competition in tournaments. Working Papers w0251, Center for Economic and Financial Research (CEFIR), URL <https://ideas.repec.org/p/cfr/cefirw/w0251.html>.
- Eisenkopf G, Teyssier S (2013) Envy and loss aversion in tournaments. *Journal of Economic Psychology* 34:240–255, URL <http://dx.doi.org/10.1016/j.joep.2012.06.006>.
- Eriksson T, Teyssier S, Villeval MC (2009) Self-selection and the efficiency of tournaments. *Economic Inquiry* 47(3):530–548, URL <http://dx.doi.org/10.1111/j.1465-7295.2007.00094.x>.
- Fransoo J, Blanco E, Mejia Argueta C, eds. (2017) *Reaching 50 million nanostores: retail distribution in emerging megacities* (CreateSpace Independent Publishing Platform), ISBN 978-1975742003.
- Freeman RB, Gelber AM (2010) Prize structure and information in tournaments: Experimental evidence. *American Economic Journal: Applied Economics* 2(1):149–64, URL <http://dx.doi.org/10.1257/app.2.1.149>.
- Fu Q, Lu J (2011) Micro foundations of multi-prize lottery contests: a perspective of noisy performance ranking. *Social Choice and Welfare* 38(3):497–517, URL <http://dx.doi.org/10.1007/s00355-011-0542-5>.
- Garcia Martinez M, Walton B (2014) The wisdom of crowds: The potential of online communities as a tool for data analysis. *Technovation* 34(4):203–214, ISSN 0166-4972, URL <http://dx.doi.org/https://doi.org/10.1016/j.technovation.2014.01.011>.
- Gareth James RTDW Trevor Hastie (2013) *An introduction to statistical learning : with applications in R* (New York : Springer, [2013] ©2013), URL <https://search.library.wisc.edu/catalog/9910207152902121>.
- Garrette B, Karnani A (2010) Challenges in marketing socially useful goods to the poor. *California Management Review* 52(4):29–47, URL https://papers.ssrn.com/so13/papers.cfm?abstract_id=1507757.
- Ge J, Honhon D, Fransoo JC, Zhao L (2021) Supplying to mom and pop: Traditional retail channel selection in megacities. *Manufacturing & Service Operations Management* 23(1):19–35, URL <http://dx.doi.org/10.1287/msom.2019.0806>.
- Gill D, Kissová Z, Lee J, Prowse V (2019) First-place loving and last-place loathing: How rank in the distribution of performance affects effort provision. *Management Science* 65(2):494–507, URL <http://dx.doi.org/10.1287/mnsc.2017.2907>.

- Gilovich T, Vallone R, Tversky A (1985) The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology* 17(3):295–314, URL [http://dx.doi.org/10.1016/0010-0285\(85\)90010-6](http://dx.doi.org/10.1016/0010-0285(85)90010-6).
- Gomes R, Shah M (2018) Last Mile Solutions for Low-Income Customers. Technical report, Shell Foundation.
- Graça SS, Barry JM, Doney PM (2016) B2B commitment building in emerging markets: the case of Brazil. *Journal of Personal Selling & Sales Management* 36(2):105–125, URL <https://doi.org/10.1080/08853134.2016.1188708>.
- Granovetter M (1985) Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91(3):481–510, URL <http://dx.doi.org/10.1086/228311>.
- Gui L, Tang CS, Yin S (2019) Improving Microretailer and Consumer Welfare in Developing Economies: Replenishment Strategies and Market Entries. *Manufacturing & Service Operations Management* 21(1):231–250, ISSN 1523-4614, 1526-5498, URL <http://dx.doi.org/10.1287/msom.2017.0700>.
- Halac M, Kartik N, Liu Q (2017) Contests for experimentation. *Journal of Political Economy* 125(5):1523–1569, URL <http://dx.doi.org/10.1086/693040>.
- Hand D, Dithrich H, Sunderji S, Nova N (2020) 2020 Annual Impact Investor Survey. Technical report, URL <https://theiiin.org/research/publication/impinv-survey-2020>.
- Hansen BB (2004) Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association* 99(467):609–618, URL <http://dx.doi.org/10.1198/016214504000000647>.
- Hansen BB, Klopfer SO (2006) Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* 15(3):609–627, URL <http://dx.doi.org/10.1198/106186006x137047>.
- Hazlett C, Xu Y (2018) Trajectory Balancing: A General Reweighting Approach to Causal Inference With Time-Series Cross-Sectional Data. *SSRN Electronic Journal* ISSN 1556-5068, URL <http://dx.doi.org/10.2139/ssrn.3214231>.
- Hilary G, Menzly L (2006) Does past success lead analysts to become overconfident? *Management Science* 52(4):489–500, URL <http://dx.doi.org/10.1287/mnsc.1050.0485>.
- Ho DE, Imai K, King G, Stuart EA (2007) Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15(3):199–236, ISSN 1047-1987, 1476-4989, URL <http://dx.doi.org/10.1093/pan/mp1013>.
- Ho DE, Imai K, King G, Stuart EA (2011) **MatchIt** : Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 42(8), ISSN 1548-7660, URL <http://dx.doi.org/10.18637/jss.v042.i08>.
- Hossain M, Kauranen I (2015) Crowdsourcing: a comprehensive literature review. *Strategic Outsourcing: An International Journal* 8(1):2–22, URL <http://dx.doi.org/10.1108/so-12-2014-0029>.
- Hoyweghen KV, Fabry A, Feyaerts H, Wade I, Maertens M (2021) Resilience of global and local value chains to the covid-19 pandemic: Survey evidence from vegetable value chains in senegal. *Agricultural Economics* 52(3):423–440, URL <http://dx.doi.org/10.1111/agec.12627>.
- Huynh D, Zuo L, Iida H (2018) An assessment of game elements in language-learning platform duolingo. *2018 4th International Conference on Computer and Information Sciences (IC-COINS)* (IEEE), URL <http://dx.doi.org/10.1109/iccoins.2018.8510568>.

- ILO (2020) Indicator description: Informal economy. Technical report, International Labor Organization, URL <https://ilostat.ilo.org/resources/concepts-and-definitions/description-informality/>.
- ILO (2021) World employment and social outlook 2021: The role of digital labour platforms in transforming the world of work. Technical report, International Labour Organization, URL <https://www.ilo.org/global/research/global-reports/weso/2021/lang--en/index.htm>.
- Imai K, Kim IS (2019) When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science* 63(2):467–490, ISSN 0092-5853, 1540-5907, URL <http://dx.doi.org/10.1111/ajps.12417>.
- Imai K, Kim IS, Wang E (2020) Matching Methods for Causal Inference with Time-Series Cross-Sectional Data. *Working Paper* URL <https://imai.fas.harvard.edu/research/tscs.html>.
- Imai K, King G, Lau O (2008a) Toward a common framework for statistical analysis and development. *Journal of Computational and Graphical Statistics* 17(4):892–913, URL <http://dx.doi.org/10.1198/106186008x384898>.
- Imai K, King G, Lau O (2008b) Toward a common framework for statistical analysis and development. *Journal of Computational Graphics and Statistics* 17:1–22.
- Imai K, King G, Lau O (2012) *Zelig: Everyone’s Statistical Software*.
- Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* 86(1):4–29, URL <http://dx.doi.org/10.1162/003465304323023651>.
- Iyer A, Palsule-Desai O (2019a) Contract design for the stockist in indian distribution networks. *Manufacturing & Service Operations Management* 21(2):398–416, URL <http://dx.doi.org/10.1287/msom.2018.0722>.
- Iyer A, Palsule-Desai O (2019b) Contract design for the stockist in indian distribution networks. *Manufacturing & Service Operations Management* 21(2):398–416, URL <http://dx.doi.org/10.1287/msom.2018.0722>.
- Jiang ZZ, Huang Y, Beil DR (2021a) The role of feedback in dynamic crowdsourcing contests: A structural empirical analysis. *Management Science* URL <http://dx.doi.org/10.1287/mnsc.2021.4140>.
- Jiang ZZ, Huang Y, Beil DR (2021b) The Role of Problem Specification in Crowdsourcing Contests for Design Problems: A Theoretical and Empirical Analysis. *Manufacturing & Service Operations Management* 23(3):637–656, ISSN 1523-4614, 1526-5498, URL <http://dx.doi.org/10.1287/msom.2020.0873>.
- Jin Y, Lee HCB, Ba S, Stallaert J (2021) Winning by learning? effect of knowledge sharing in crowdsourcing contests. *Information Systems Research* 32(3):836–859, URL <http://dx.doi.org/10.1287/isre.2020.0982>.
- Jue DM (2015) Technology for development: Hammers looking for nails? URL <http://archive.skoll.org/2015/09/16/technology-for-development-hammers-looking-for-nails/>.
- Kalra A, Shi M (2001) Designing optimal sales contests: A theoretical perspective. *Marketing Science* 20(2):170–193, URL <http://dx.doi.org/10.1287/mksc.20.2.170.10193>.
- King G, Tomz M, Wittenberg J (2000) Making the Most of Statistical Analyses: Improving Interpretation and Presentation. *American Journal of Political Science* 44(2):347, ISSN 00925853, URL <http://dx.doi.org/10.2307/2669316>.
- Kong Q, Granic GD, Lambert NS, Teo CP (2020) Judgment error in lottery play: When the hot

- hand meets the gambler's fallacy. *Management Science* 66(2):844–862, URL <http://dx.doi.org/10.1287/mnsc.2018.3233>.
- Künzel SR, Sekhon JS, Bickel PJ, Yu B (2019) Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116(10):4156–4165, ISSN 0027-8424, URL <http://dx.doi.org/10.1073/pnas.1804597116>.
- laElena Asparouhova, Hertzel M, Lemmon M (2009) Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers. *Management Science* 55(11):1766–1782, URL <http://dx.doi.org/10.1287/mnsc.1090.1059>.
- Lazear EP, Rosen S (1981) Rank-order tournaments as optimum labor contracts. *Journal of Political Economy* 89(5):841–864, URL <http://dx.doi.org/10.1086/261010>.
- Levy JS (1996) Loss aversion, framing, and bargaining: The implications of prospect theory for international conflict. *International Political Science Review / Revue internationale de science politique* 17(2):179–195, ISSN 01925121, URL <http://www.jstor.org/stable/1601302>.
- Liker JK, Choi TY (2004) Building deep supplier relationships. *Harvard Business Review* 82(12):104–113, URL <https://hbr.org/2004/12/building-deep-supplier-relationships>.
- Lim N, Ahearne MJ, Ham SH (2009) Designing sales contests: Does the prize structure matter? *Journal of Marketing Research* 46(3):356–371, URL <http://dx.doi.org/10.1509/jmkr.46.3.356>.
- List JA, van Soest D, Stoop J, Zhou H (2020) On the role of group size in tournaments: Theory and evidence from laboratory and field experiments. *Management Science* 66(10):4359–4377, URL <http://dx.doi.org/10.1287/mnsc.2019.3441>.
- Liu L, Wang Y, Xu Y (2020) A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data. *SSRN Electronic Journal* ISSN 1556-5068, URL <http://dx.doi.org/10.2139/ssrn.3555463>.
- Ludwig S, Lünser GK (2012) Observing your competitor – the role of effort information in two-stage tournaments. *Journal of Economic Psychology* 33(1):166–182, URL <http://dx.doi.org/10.1016/j.joep.2011.09.011>.
- Mason W, Watts DJ (2010) Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.* 11(2):100–108, ISSN 1931-0145, URL <http://dx.doi.org/10.1145/1809400.1809422>.
- Moldovanu B, Sela A (2001) The optimal allocation of prizes in contests. *American Economic Review* 91(3):542–558, URL <http://dx.doi.org/10.1257/aer.91.3.542>.
- Morgan SL, Winship C (2015) *Counterfactuals and causal inference: methods and principles for social research*. ISBN 978-1-107-58799-1, URL <https://doi.org/10.1017/CB09781107587991>, oCLC: 898207346.
- Myerson R, Wärneryd K (2006) Population uncertainty in contests. *Economic Theory* 27(2):469–474, URL <http://dx.doi.org/10.1007/s00199-004-0605-2>.
- Nozaki Y (2018) Local Kirana Shops Offer Various Business Opportunities in Retail and Distribution Market in India. Technical report, Mitsui & Co. Global Strategic Studies Institute.
- OECD, ILO (2019) *Tackling Vulnerability in the Informal Economy* (OECD), URL <http://dx.doi.org/10.1787/939b7bcd-en>.
- Rabin M (2002) Inference by believers in the law of small numbers. *The Quarterly Journal of Economics* 117(3):775–816, URL <http://dx.doi.org/10.1162/003355302760193896>.

- Revoredo-Giha C, Renwick A (2016) Market structure and coherence of international cooperation: the case of the dairy sector in malawi. *Agricultural and Food Economics* 4(1), URL <http://dx.doi.org/10.1186/s40100-016-0052-y>.
- Segev E (2020) Crowdsourcing contests. *European Journal of Operational Research* 281(2):241–255, URL <http://dx.doi.org/10.1016/j.ejor.2019.02.057>.
- Shao B, Shi L, Xu B, Liu L (2012) Factors affecting participation of solvers in crowdsourcing: an empirical study from china. *Electronic Markets* 22(2):73–82, URL <http://dx.doi.org/10.1007/s12525-012-0093-3>.
- Shukla S, Bairiganjan S (2011) The base of pyramid distribution challenge: evaluating alternate distribution models of energy products for rural base of pyramid in india. Technical report, Centre for Development Finance, Institute for Financial and Management Research.
- Shupp R, Sheremeta RM, Schmidt D, Walker J (2013) Resource allocation contests: Experimental evidence. *Journal of Economic Psychology* 39:257–267, URL <http://dx.doi.org/10.1016/j.joep.2013.09.001>.
- Simanis E (2012) Reality Check at the Bottom of the Pyramid. *Harvard Business Review* URL <https://hbr.org/2012/06/reality-check-at-the-bottom-of-the-pyramid>.
- Sinha P, Kumar S, Prakash S (2020) Measuring and mitigating the effects of cost disturbance propagation in multi-echelon apparel supply chains. *European Journal of Operational Research* 282(1):148–160, URL <http://dx.doi.org/10.1016/j.ejor.2019.09.015>.
- Stuart EA, Green KM (2008) Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology* 44(2):395–406, URL <http://dx.doi.org/10.1037/0012-1649.44.2.395>.
- Szymanski S, Valletti TM (2005) Incentive effects of second prizes. *European Journal of Political Economy* 21(2):467–481, URL <http://dx.doi.org/10.1016/j.ejpolco.2004.07.002>.
- Tajedin H, Madhok A, Keyhani M (2019) A theory of digital firm-designed markets: Defying knowledge constraints with crowds and marketplaces. *Strategy Science* 4(4):323–342, URL <http://dx.doi.org/10.1287/stsc.2019.0092>.
- Tuori MA (2012) Strengthening informal supply chains : the case of recycling in bandung, indonesia.
- Uzzi B (1997) Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly* 42(1):35–67, URL <http://dx.doi.org/10.2307/2393808>.
- Viswanathan M, Rosa JA, Ruth JA (2010) Exchanges in marketing systems: the case of subsistence consumer- merchants in Chennai, India. *Journal of Marketing* 74(3):1–17.
- Wang J, Ipeirotis PG, Provost F (2017) Cost-effective quality assurance in crowd labeling. *Information Systems Research* 28(1):137–158, URL <http://dx.doi.org/10.1287/isre.2016.0661>.
- Wang Y, Cai H, Li C, Jiang Z, Wang L, Song J, Xia J (2013) Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study. *PLoS ONE* 8(12):e81045, ISSN 1932-6203, URL <http://dx.doi.org/10.1371/journal.pone.0081045>.
- WB (2021) The long shadow of informality: Challenges and policies. Technical report, The World Bank, URL <https://www.worldbank.org/en/research/publication/informal-economy>.
- Webb JW, Kistruck GM, Ireland RD, Ketchen Jr DJ (2010) The entrepreneurship process in base of the pyramid markets: The case of multinational enterprise/nongovernment organization alliances. *Entrepreneurship Theory and Practice* 34(3):555–581, URL <http://dx.doi.org/10.1111/j.1540-6520.2009.00349.x>.

Zhang S, Singh PV, Ghose A (2019) A Structural Analysis of the Role of Superstars in Crowdsourcing Contests. *Information Systems Research* 30(1):15–33, ISSN 1047-7047, 1526-5536, URL <http://dx.doi.org/10.1287/isre.2017.0767>.