Empirical essays on information asymmetries on digital platforms

Chiara Belletti

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAT017

Thèse de doctorat

Empirical Essays on Information Asymmetries on Digital Platforms

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Sciences Économiques

Thèse présentée et soutenue à Palaiseau, le 24 Mai 2024, par

CHIARA BELLETTI

Composition du Jury :

Marc Bourreau Professeur d'Économie, Télécom Paris	Président/Examinateur
Jose De Sousa Professeur d'Économie, Université Paris-Panthéon-Assas	Rapporteur
Tobias Kretschmer Professeur de Gestion, Ludwig Maximilian University Munich	Rapporteur
Ulrich Laitenberger Professeur d'Économie, Télécom Paris & Tilburg University	Directeur de thèse

Acknowledgements

I extend my gratitude to all those who have played a significant role, both directly and indirectly, in the completion of this thesis.

First and foremost, I express my appreciation to my supervisor, *Ulrich Laitenberger*, for his invaluable guidance, support, and constant availability throughout this journey.

My deepest thanks go to the members of my committee, *Marc Bourreau*, *Jose De Sousa*, and *Tobias Kretschmer*, for generously dedicating their time to evaluate the quality of my work.

I am immensely indebted to my exceptionally smart and kind coauthors, *Louis Pape* and *Michelangelo Rossi*. Their guidance has been invaluable, and I cannot adequately express my gratitude for all they have taught me.

I am deeply grateful to all the faculty members of our department for their role in shaping me as a scholar and for providing enriching experiences during the past years. In particular, I thank *Marc Bourreau* for his kindness, support and the valuable career advice he offered to me. My gratitude also extends to *David Bounie*, *Laurie Ciaramella*, *Lukasz Grzybowsky*, *José-Luis Moraga-Gonzáles*, and *Patrick Waelbroeck*.

I would like to thank *Christine Zulehner* and *Matthias Hirth* for evaluating my progress and for their feedback during my mid-term defense.

Special acknowledgment is due to *Paola Tubaro* and *Antonio Casilli* for supporting my thesis. The insightful discussions and expertise shared during Diplab meetings have profoundly enriched my understanding and perspective over platform labor. I am also grateful to *Marie-Laure Allain* for her support in extending my thesis funding and for the opportunity to join the CREST team.

A heartfelt thanks to *Tobias Kretschmer* and the entire ISTO team for their warm welcome during my visit in Munich in 2023 and for the inestimable opportunity to join their team. Special appreciation goes also to *Irene Bertschek* and her team at ZEW in Mannheim for the enriching discussions during my visit in 2021.

I also acknowledged *Nhatvi Nguyen* for providing feedback, data and computing ca-

pability and *Alexander Kanunov* for providing technical support.

I have been fortunate to share the doctoral experience with exceptional colleagues who became precious friends to me, in particular *Guillaume Thébaudin*, *Pierre-François Darlas* and *Chloé Breton*. I am also grateful for the time spent with all the other PhD students and the visitors of the department: *Raphaela Andres*, *Joshua Brand*, *Martin Delville*, *Quentin Durand*, *Leonardo Leone*, *Angela Munoz*, *Federico Navarra*, *Luca Rossi*, *Felix Shleef* and *Ivan Shchapov* .

Lastly, and most importantly, I could not have embarked and pursued on this journey without the steadfast support of my friends and family.

In particular, I extend my gratitude to *Caroline*, who has been a fundamental presence in my Parisian life, and my friends *Camilla, Francesca, Ottavia* and *Stefania*.

I would not be where I am today without the incredible support of my loving and supporting parents, *Annamaria* and *Luciano* who have always believed in me, and to whom I dedicate this milestone in my life.

Finally, this thesis is also dedicated to *Riccardo*, with whom I have not only shared this PhD experience but also my everyday life in the past three years. I can never be grateful enough for his support, encouragement, and love.

Contents

0 Introduction

0.1 Digital Platforms and Asymmetry of Information

Over the past decades, digital platforms, such as Amazon Marketplace, Facebook, Uber, Upwork and Tinder, have changed multiple aspects of our daily lives, from the way we shop, connect, move, and even work and date. Defined by Belleflamme and Peitz (2021), platforms are entities that bring together economic agents and actively manage "network effects" among them. These effects involve the impact an additional participant has on others economic agents within the market. Platforms oversee different types of network effects, including those within the same user group (i.e. "within-group" externalities) and those that transcend diverse groups on the market (i.e. "cross-group" externalities). The latter occurs when the utilities of two distinct groups on the platform influence each others (Rochet and Tirole 2003) and are leveraged by so called "two-sided" platforms such as Amazon, which connects sellers and buyers, or Airbnb, which facilitates interactions between hosts and guests in the short-term accommodation industry.

Amazon and Airbnb are just two prominent examples of the various digital platforms that have emerged due to the internet's enabling capabilities, which provide tools and technologies facilitating managerial and coordination functions for intermediary firms. Digital platforms use such tools to regulate sellers' entry into the market and establish quality standards (Evans 2020), to allow or prevent compatibility with other platforms (Crémer, Rey, and Tirole 2000; Jullien and Sand-Zantman 2021) and even to shape the level of freedom economic agents have over their interactions (Liu, Yildirim, and Zhang 2021). Another crucial aspect is the control that platforms wield over information disclosure. Digital platforms leverage data from agents interactions for different purposes. First, to build recommendation systems. Platforms indeed exert influence on their users' choices through search rankings, search filters and personalized recommendation systems. A notable example is Amazon Marketplace, where the platform positions products based on a recommendation algorithm, affecting buyers' decisions and shaping their purchasing behavior. Second, platforms use transactions data to implement rating systems based

on users' feedback. Feedback typically includes numerical ratings and written comments, providing a way for individuals to share their experiences and express satisfaction or dissatisfaction about the interactions with others agents on the platform.[1]

Ratings and reviews mitigate a common market failure: "information asymmetry", namely a situation where different sides of the platform lack access to the same information. Usually, this translates into buyers having less information than sellers about the quality of goods or services offered for sale (Belleflamme and Peitz 2021). Information asymmetry is associated with two class of risks: "adverse selection" and "moral hazard". Adverse selection refers to a scenario where one side of the economic transaction possesses more information than the others, leading to potentially unfavorable outcomes. Akerlof (1970) introduced this concept using the example of the used car market. In such market, buyers lack tools such as certifications or warranties to objectively assess products' quality and rely solely on prices to infer a vehicle's conditions. This difficulty in distinguishing between vehicles can lead to the dominance of "lemons" (i.e. low-quality products) and the exclusion of higher-quality cars from the market. The economic concept of adverse selection primarily addresses the inherent, or fixed component, of service quality. However, the overall quality of a transaction is also influenced by the level of attention, effort, and care that parties invest in the transaction process. Moral hazard refers to a situation in which one party, typically after entering into an agreement or a contract, may be inclined to deviate from it and act in a way that could be detrimental to the interests of the other market side, typically after the financial transfer occurred. Moral hazard can manifest in various contexts, including insurance, financial markets, corporate governance, and principal-agent relationships (Holmström 1979). For example, in the insurance industry, moral hazard occurs when policy-holders are more likely to engage in risky behavior because they know they are protected by insurance coverage.[2]

1. Some two-sided platforms even enable both sides of the market to mutually review each other (e.g., Airbnb, Uber, Upwork). Finally, on certain platforms, the evaluation of quality and agents' behavior extends beyond participants on the platform to review offline commercial activities such as restaurants and businesses. For instance, Yelp serves as a notable example of a platform that provides reputation systems for offline services.

2. Moral hazard and optimal health insurance coverage is discussed in the seminal paper by Pauly 1968.

On digital platforms information asymmetry is exacerbated by anonymity and geographical distance among economic agents and adverse selection and moral hazard tend to coexist. To illustrate this scenario, consider a platform for crowdfunding where entrepreneurs seek funds for their projects. Due to the anonymity of the platform, investors may lack complete information about the project's quality, feasibility, and the entrepreneur's skills. This lack of information can result in adverse selection, where projects that are oversold receive funding. Simultaneously, once funding is secured and the project is underway, the physical distance and absence of face-to-face interactions can give rise to moral hazard. The entrepreneur may be tempted to deviate from the initially promised project plan, misallocate funds, or even abandon the project. Dellarocas (2006) describes two mechanisms for how reputation systems help mitigating information asymmetry. First, ratings and reviews act as a "signaling device", assisting buyers in assessing the quality of sellers and thereby reducing adverse selection. Second, ratings and reviews function as a "monitoring device" and play a sanctioning role for the sellers, addressing issues related to moral hazard. Indeed, in a single interaction, agents might not have strong incentives to behave optimally. However, through repeated interactions, reputation serves as a record of performance and behaviors. This means that misconduct carries repercussions for future interactions, incentivizing agents to maintain a positive reputation over time.

0.2 Impact of Reputation Systems on Digital Platforms

Numerous studies have delved into both the direct and indirect consequences of reputation systems on digital platforms.[3] First, the literature describes how reputation systems function as a mechanism for incentivizing and improving quality provision. For instance, Farronato and Zervas (2022) provide empirical evidence that restaurants improve hygiene when they are more exposed to review platforms. Second, online reviews are shown have positive impact on sellers' sales and revenue (Chevalier and Mayzlin 2006; Cabral and Hortacsu 2010; Luca 2011; Anderson and Magruder 2012 and Yoganarasimhan 2013).

3. Refer to Rossi (2018) for an in-depth literature review on this topic.

For instance, Chevalier and Mayzlin (2006) investigate the impact of reputation systems on book sales at two online book retailers, Amazon and Barnes & Noble and demonstrate that enhancements in a book's reviews result in a notable increase in sales of that book on the e-store.

Moreover, ratings systems contributed to increased market transparency and enabled the flourishing of otherwise marginal markets. One notable example is the "sharing economy", which includes sectors such as short-term housing and car rides sharing. In the past, the idea of accepting a ride from a stranger would have been considered with caution by many, nowadays, numerous individuals use regularly various apps that facilitate ride-sharing (e.g., BlaBlaCar, Uber, and Lyft). Rating systems provide a mechanism for users to share their experiences, cultivating a sense of reliability and confidence that motivates participation in these markets and contributes to their growth. Online reputation also had a significant role in transforming the so-called "gig economy" (Woodcock and Graham 2019). In the past, gig opportunities relied heavily on word-of-mouth recommendations. Now, the gig work market has significantly shifted online, allowing to hire peers for various tasks, such as assembling furniture (e.g., TaskRabbit) or taking piano classes (e.g., Superprof) after having checked the reviews left by previous clients. Furthermore, geographical distance in labor market may be overcome, as reputation systems on freelance platforms (e.g., Upwork and Guru) enable the collection of information necessary to hire a professional, even when geographically distant.[4]

However, ratings and reviews may not always be fully efficient. To mitigate adverse selection, they are valuable only if they provide relevant and accurate information. However several sources of bias hinders the informativeness of ratings and reviews. Belleflamme and Peitz (2021) summarize several sources of noise in the reputation signal: bad understanding (i.e. feedback focuses on irrelevant - to the other users - details of the transaction), idiosyncratic tastes of reviewers and unexpected events beyond the control of sellers (e.g., a delayed delivery due to road congestion). In digital markets, the ease

4. Freelance platforms serve as online platforms that link individuals seeking services that can be provided electronically with freelancers who offer their services on a per-project basis or for a fixed hourly rate. These platforms incorporate a simple system enabling the assessment of feedback and past client experiences.

of reputation manipulation through fake reviews can lead to biased and inaccurate representations of product or service quality (Mayzlin, Dover, and Chevalier 2014; Zervas, Proserpio, and Byers 2021). Bias and reputation "inflation" (i.e. left-skewed distribution of ratings) may also arise from "herding behavior", where users follow the mass in their evaluations (Muchnik, Aral, and Taylor 2013) or from social reciprocity and fear of retaliation in bilateral review systems (Klein, Lambertz, and Stahl 2016; Fradkin, Grewal, and Holtz 2018 and Zervas, Proserpio, and Byers 2021). Furthermore, the voluntary nature of providing ratings and reviews and the associated costs for the user, may result in a scenario where only users with exceptionally positive or negative experiences choose to submit reviews, resulting in a J-shaped distribution (Mayzlin, Dover, and Chevalier 2014).

Another class of issues related to ratings and reviews on digital platforms pertains to the effect of reputation systems on market structure. Reputation systems create network effects, wherein the value of reviews increases with the platform's user base. Since these effects are platform-specific and reputation transferability is rare, there is a risk of a winner-take-most scenario, where a significant portion of users gravitate towards a single platform to access more extensive review information (Belleflamme and Peitz 2018; 2021).

0.3 Thesis' Contribution

At the state of art, there are still some open questions concerning the behaviour of economic agents in asymmetry of information and the efficiency of reputation on digital platforms. This thesis address two issues that have received limited attention thus far. The first pertains to the "end-game" effect. Often, ratings and reviews are designed to function within an infinite game. However, it remains unclear how they operate at the end of a seller's career on a digital platform. Literature on "career concerns" in offline markets include theoretical (Holmström 1999) and empirical evidence of a decline in effort at the end of a worker career (Gibbons and Murphy (1992)) or in analogous situations where reputation concerns stop to be at play (Miklós-Thal and Ullrich (2016)). Yet, there is scarce empirical evidence demonstrating that the same mechanisms apply in online

markets. Cabral and Hortacsu (2010) find that ratings are lower in the last transactions of sellers on eBay. However, they do not determine whether this effect is due to end-game concerns affecting sellers' effort or if, reversely, sellers' exit is driven by the collection of poor ratings, prompting sellers to leave the unprofitable platform. On a different note, Xu, Nian, and Cabral (2020) explore non-pecuniary incentives on the Q&A platform Stack Overflow. Their study reveal that users contribute content to signal their quality to potential employers and that they decrease activity on the platform after securing a new job. This suggests that new job opportunities reduce reputation concerns related to users' careers on the platform. In their study, the authors do not rule out the possibility of workers returning to the platform and restarting their incentives if they seek another job. However, little is known about the causal relationship on sellers' effort associated with a definitive career termination. The first chapter of this thesis aims at answering this question by investigating how "end-game" concerns influence sellers' decisions in the short-term rental market.

A second open question has to do with the effect of the information asymmetry in digital markets where standard reputation tools, such as ratings and reviews, are insufficient in providing informative reputations measures, as in the case of online labor markets for "crowd-sourcing" of "micro-tasksing". Micro-tasking platforms are online marketplaces that connect outsourcing firms with a large and diverse crowd of contributors for completion of small and simple tasks under a piece-rate compensation. Tasks can range from data entry and annotation to generation of web traffic or testing new apps. Important platforms include Amazon Mechanical Turk and microWorkers. On micro-tasking platforms, workers usually self-select into tasks and there is minimal ex-ante screening by the firm of which workers can access and execute their job listings. Additionally, limited interactions occur between employer and employee, and the modest financial incentives might lead to a sub-optimal situation where employers overlook quality checks and workers, consequently, shirk. As workers' reputation on crowd-sourcing platforms is frequently linked to rejection rates, this contributes to an inflated and inadequately informative reputation system. The first and second chapters of this thesis study how users behave on

these platforms and explore alternative methods to elicit quality. In particular, the second chapter investigates the behavior of firms outsourcing AI related tasks to ensure quality of execution by task's design, incentives, workers selection and, finally, quality check. The third chapter assesses the effective quality of crowd-sourced data annotation tasks and tests the role of alternative schemes that a micro-tasking platform can implement to enhance overall quality of work. The remaining part of this introduction summarizes the three chapters of this thesis.

Chapter 1: Reputation Concerns and the End-Game Effect: When Reputation Works and When it Does Not.

The first chapter of this thesis, co-authored with Elizaveta Pronkina and Michelangelo Rossi, explores how end-game considerations, namely the anticipation of a forthcoming exit from a marketplace, influence sellers' effort decisions on digital platforms.

Understanding whether the effectiveness of reputation systems diminishes in the final transactions of sellers is crucial in digital markets, where low entry and exit costs may result into a high turnover of sellers who have brief engagements on the platform and therefore being less exposed on reputation incentives.

To address this question, we collected data from the short-term rental platform Airbnb and examined the impact of an exogenous source of anticipation of hosts' exit from the marketplace on host's effort. We measure the latter via listings' ratings in effort-related categories (i.e. check-in, cleanliness, and communication). To avoid confounding the end-game effect with endogenous exit triggered by collection of bad ratings, we employ a quasi-natural experiment. We leverage a regulation on short-term rentals adopted by the City of Los Angeles at the end of 2018 to identify hosts who anticipated their imminent exit from the platform due to non-compliance with new eligibility rules (orthogonal to the cumulative host's reputation). Employing a Difference-in-Differences methodology and an Event Study, we examine how effort-related ratings of listings leaving the platform due to the regulation, evolved, compared to ratings on location, following the regulatory announcement and during its implementation. We choose ratings on location as control group since location represents a fixed quality dimension of the listing and, thus, it is

unrelated to host's effort.

We document a statistically significant decrease in effort-related ratings during the last periods of the listing's presence of the platform. In the months of the policy implementation, effort related ratings decreased by around 1.5% to 2.2% of the mean value. Our findings underscore the adverse effects of end-game considerations on the power of reputation systems as a tool to mitigate moral hazard at the end of sellers' careers. Even though outside the scope of the paper, these results offer valuable insights for platform managers designing reputation systems and prompts an open discussion on the optimal length of reputation and on the adoption of alternative incentive tools.

Chapter 2: Crowd-sourcing AI Related Tasks: Insights from an Online Labor Platform

The second chapter, joint work with Ulrich Laitenberger and Paola Tubaro, provides new descriptive evidence on how crowd-sourcing platforms are used to outsource AI related jobs, mostly data training.

Crowd-sourcing platforms offer a scalable and cost-effective solution for outsourcing the collection of the human inputs required for AI models' data training. However, specific characteristics of platform labor may raise ethical and privacy-related concerns. In addition, the presence of numerous workers contributing to the same project complicates the selection of an adequately skilled labor force and adds complexity to quality investigations. This introduces the risk of gathering poorly executed work and using misannotated data for training AI models.

Deriving insights from proprietary data from a leading commercial crowd-sourcing platform, the chapter studies the the demand volume and the content of AI related jobs outsourced on the platform. It also investigates the behaviors adopted by requesters to ensure the quality of tasks execution. To identify AI related tasks, we employ a text analysis approach, detecting keywords associated with data annotation and generation in the jobs' title and description. We complement this approach with a more straightforward identification based on the labels chosen by requesters to categorize tasks. First, the paper emphasizes a growing demand for data-work on the platform since 2019. We then

employ a text mining approach to gather keywords and cluster them based on actions and industry scope, aiming to comprehend the demand. We find that data collection tasks primarily serve market research purposes (e.g., collection of products' prices). In more than half of tasks about data collection and generation, workers are required to generate first-hand information, either by recording themselves, taking pictures or answering to certain questionnaires. Data annotation tasks often require the identification of emotions and spatial objects for training AI models. Industries dealing with sensitive data are still relatively marginal on the platform, highlighting existence of some requesters' privacy concerns in sharing sensitive data with the "crowd".

Finally, we use a regression framework to identify specific features that differentiate the demand for AI related work from other jobs. While we do not observe significant changes in monetary incentive, we show that requesters of AI related jobs make significantly more use of ex-ante selection of workers, particularly through predefined groups of workers or geo-targeted demand. Higher probability of tasks' rejection in data annotation highlights the greater relevance of quality in this domain and a more substantial effort by requesters to monitor executed tasks. In exploring these different strategic dimensions of the requesters' behaviour, we offer valuable insights for new outsourcing firms, elucidating the methods they can employ to ensure collection of quality output. Furthermore, we advise the platform on the prevalent tools favored by their clients, which can be strengthened to enhance its attractiveness.

Chapter 3: Moral Hazard in Micro-Tasking. Evidence from a Structural Model

This chapter, coauthored with Louis Daniel Pape, evaluates quality of the data annotation tasks outsourced on a leading commercial micro-working platform and explore the effects of monetary incentives.

Crowd-sourcing platform provide data used to train machine learning algorithms and artificial intelligence. However, a classical Principal-Agent problem, fostered by low monetary rewards of outsourced tasks, limits the quality of the data produced on such platforms. This problem results from firms not monitoring the quality of the work done with

sufficient frequency. Quality of executed tasks remains unobserved in the platform data, while we observe if each task has been validated (and paid) or rejected by the outsourcing firm. However, validation is limited in informativeness when stemming from a lack of investigation by the firm.

To disentangle the mechanisms governing tasks rejection, namely the worker effort and the firm investigation decision, we adopt a structural approach, modelling the simultaneous demand and supply of effort on the platform. The model considers the moderating impact of expectations from each platform's side on the other side's choice. The value of wages for the worker is influenced by the expected investigation they will undergo. Similarly firms take into account the expected effort by the worker when deciding if monitoring quality of tasks. The equilibrium outcome, observed as rejection/validation decisions in the data, is derived through fulfilled rational expectations.

We estimate our model with propriety data from a leading micro-tasking platform and we reveal that rejection rates underestimate quality of executed tasks. Additionally, we simulate different counter-factual incentive schemes to induce higher quality work. In partial equilibrium, we find that a wage penalty for workers with a rejected task could induce higher effort and require less monitoring from the firms. This strategy would provide additional revenue to the platform which she could then distribute to the outsourcing firms as a subsidy to encourage them to monitor (and then reject) tasks.

Our study provides platforms with a tangible measure to assess work quality through the observed data of rejection and investigation, offering valuable insights for informed managerial decisions and showing the role of monetary incentives in enhancing task quality and mitigating the risk of introducing undetected biases in the final work applications.

1 Reputation Concerns and the End-Game Effect: When Reputation Works and When it Does Not

This paper is written together with Elizaveta Pronkina (Université Paris-Dauphine - PSL) and Michelangelo Rossi (Télécom Paris, Institut Polytechnique de Paris).[5]

1.1 Introduction

Digital platforms implement reputation systems to reduce information asymmetries between users and provide sellers with incentives to exert effort over time to build and maintain their reputation. Positive consumers feedback is key to prosper on the platform (Cabral and Hortacsu 2010; Luca 2011; Anderson and Magruder 2012). Yet, career concerns and the power of reputation incentives may wane when future periods to benefits from reputation are limited (Gibbons and Murphy 1992, Holmström 1999). This is a severe risk for digital markets, where relatively low costs of entry and exit may lead to a high turnover of sellers who spend short time on platforms. However, limited evidence of moral hazard at the end of sellers' career on digital platforms have been produced.

To fill this gap, this paper investigates how end-game considerations, namely the anticipation of a forthcoming exit from the platform, affect sellers' effort decisions in their last transactions. Our analysis focuses on the behavior of hosts on the platform Airbnb. The market for short-term rentals (STR) has been increasingly regulated worldwide and many hosts have been compelled to leave the market due to their inability to comply with stricter eligibility criteria and licensing requirements. The anticipated announcement of these policies before their effective enforcement means that hosts are aware of their imminent exit from the marketplace and can adjust their behavior accordingly. This

5. This work was supported by the French Research Agency (ARN) under grant ANR-19-CE10-0012 ("HUSH"). We would like to extend our gratitude to Anahid Bauer, Jörg Claussen, Laszlo Goerke, Tobias Kretschmer, Ulrich Laitenberger, Mark J. Tremblay, and Nikhil Vellodi. Helpful feedback was received at the seminars at Télécom Paris; the 3rd Crowdworking Symposium 2021; the AFREN doctoral workshop 2022; the ZEW ICT conference 2022; the CESifo area conference on Economics of Digitization 2022; the 6th Doctoral Workshop on the Economics of Digitization 2023; the Digital Economy Workshop in Rotterdam 2024 and the 15th Paris Digital Economics Conference. All errors are ours. E. Pronkina acknowledges that this work was done prior to the author joining Amazon.

could undermine the efficiency of platforms' reputation systems if reviews and ratings become less effective, as hosts approaching the end of their career may attempt to "milk" their accumulated reputation and misbehave.

To empirically evaluate how end-game considerations affect effort choices of hosts, we collect scraped data from Airbnb in Los Angeles. We measure effort from the listings' evaluations by guests in rating categories such as check-in, communication and cleanliness. To identify anticipated hosts' exit from the platform, we take advantage of the implementation of the Home-Sharing Ordinance (HSO) in the City of Los Angeles. The HSO was approved by the city council of Los Angeles in December 2018 as a tool to regulate the short-term rental market in the city. According to the regulation, hosts willing to rent out their dwellings for less than 30 consecutive days, had to register and pay a license fee. The eligibility for getting licensed was conditional on a set of dwelling's characteristics but mostly it was limited to the host's primary residence. STR was limited to maximum 120 days in a calendar year. Airbnb engaged in the regulation enforcement by removing from the marketplace those listings who failed to provide the license number within the due period. Implementation of the HSO lasted four months, from the beginning of July to the end of October 2019, during which more than one fifth of short-term listings in the City of Los Angeles left the platform. Due to the lag between policy approval (and its announcement) and policy implementation, ineligible listings, or those unwilling to register could anticipate their forthcoming exit from the platform.

The eligibility criteria set by the regulation were unrelated to the rating history of listings. Thus, restricting the analysis to hosts renting in the City of Los Angeles that exited due to the HSO, allow us also to disentangle the effect of end-game considerations from exits induced by the accumulation of negative evaluations. In a Difference-in-Differences (DiD) setting and with an Event Study (ES) analysis, we compare the evolution, before and after the HSO, of ratings in effort-related categories with ratings on the listing's location. We choose ratings on location as control group since location is independent to the host's effort and therefore not affected by the regulation.

This paper contributes to different strands of the literature. First, it adds to the

economic literature exploring of how reviews motivate sellers' efforts and improve quality differently according to the moment of their career. Fan, Ju, and Xiao (2016) analyze how sellers manage their reputation through the life cycle in the Chinese platform Taobao. They show that new and experienced sellers manage their reputation over time in different ways. New sellers do not increase prices after receiving the first positive reviews but, they keep them low to further boost their volumes of trade. After many reviewed transactions, new sellers become experienced sellers. Thus, with a stronger reputation, they exploit their position to increase prices. Prices are not the only variable affecting buyers' value of a transaction. In almost all digital platforms, sellers can affect the quality of the service over time through effort. Cabral and Hortacsu (2010) report that, on eBay, after the first negative rating, further negative feedback follows 25 percent more frequently. Still, these new negative reviews have a lower impact on the sellers' performance. With a high reputation, the incentives to behave well are also high. Conversely, if the level of reputation is low because of a negative review, sellers are less motivated to perform well. They also find that ratings in the last transactions of sellers is lower than their average score. However, the authors do not determine whether this effect is due to end-game concerns affecting sellers' effort or if, reversely, sellers' exit is driven by the collection of poor ratings, prompting sellers to leave the unprofitable platform.

To show how future career prospects affect effort incentives, Miklós-Thal and Ullrich (2016) study career concerns of professional soccer players during the selections for national teams for the European Cup. The authors observe that players with some chances to be selected for the national teams perform better during the selection period. This effect is not present for players who cannot be selected for external reasons. Xu, Nian, and Cabral (2020) explore non-pecuniary incentives for users on the Q&A platform Stack Overflow. They find that users provide content as a way to signal their quality to potential future employers. Users accumulate less positive feedback and reduce their activity on the platform right after finding a new job. Thus, the new job opportunity reduces reputation concerns related to users' careers on the platform. In their study, the authors do not rule out the possibility of workers returning to the platform and restarting their

incentives if they seek another job. Differently, our paper focus of causal relationship on sellers' effort associated with a definitive career termination on a digital platform.

By focusing on potential weaknesses of reputation systems at the moment of sellers' exit from an online marketplace, we also contribute to the stream of literature that studies limitations of ratings and reviews. For instance, extensive evidences have been produced on the bias of reputation systems in measuring effective quality of products or services (Dellarocas and Wood 2008; Mayzlin, Dover, and Chevalier 2014; Zervas, Proserpio, and Byers 2021). However, little has been said on the possibility that ratings and reviews may fail in providing proper incentives to sellers when they can anticipate their exit from an online marketplace. To the best of our knowledge, we are the first to quantify this effect with a shock-based identification strategy.

Finally, we contribute to two academic discourses on Airbnb STR market. On one side, we provide specific insights into the growing body of literature examining rating systems on the platform (Fradkin, Grewal, and Holtz 2021; Zervas, Proserpio, and Byers 2021; Carnehl et al. 2022 and Rossi 2023). On the other, our study aligns with the literature on the consequences of STR regulations. While our analysis doesn't directly estimate the impact of such regulation, we leverage it in our identification strategy. Notable works in this context include Koster, Van Ommeren, and Volkhausen (2021), who investigate Airbnb influence on housing prices in the county of Los Angeles. Li, Kim, and Srinivasan (2022) take a structural approach and estimate the effects of potential regulations of STR on the housing rental market. Finally, Bekkerman et al. (2023) estimate the effect of STR on residential investment.

Our analysis documents that, when hosts anticipate exit, effort-related ratings decrease significantly in their last transactions. The maximum magnitude of the effect varies from about -0.07 points for check-in, -0.08 points for communication and up to -0.10 points for cleanliness ratings. This corresponds to around 1.5 to 2.2% decrease from the pre-regulation average rating. The size of the DiD estimates is not of a negligible magnitude considering that ratings on Airbnb (in a scale from 1 to 5 stars) are particularly sticky and left-skewed. While the end-game effect does not seem to be influenced

by listings' ratings on the overall experience and location before regulation, the decline in ratings for check-in and communication during the final transactions is slightly more pronounced for "professional" hosts. This suggests that the inherent value of participating in the sharing economy for non-professional sellers could partially mitigate moral hazard at the end of their career.

To validate our result, we exploit the geographical variation of the policy. The HSO indeed only applied to the City of Los Angeles. The other 87 cities and unincorporated areas in the county of Los Angeles were not subject to the regulation nor to simultaneous similar policies.[6] We perform a DiD analysis comparing effort-related ratings of listings in the City of Los Angeles who exited during the HSO implementation, with effort-related ratings of listings located in other cities in the county (who left the platform in the same period). Similarly to out main specification, we document that listings affected by the HSO experienced a statistically significant decrease in ratings about check-in and communication during the regulation implementation. Ratings about location are not affected by the regulation, confirming the validity of our control group in the main specification.

Our results suggest that when hosts expect to stay on the platform only for a few more transactions, the power of reputation incentives to reduce moral hazard weakens. Platform operators may want to consider this effect while designing their ratings and reviews system, especially when faced with high turnover or anticipating periods of significant sellers' exit, such as those resulting from external regulations.

The remaining part of the paper is organized as follows: Section 1.2 describes the empirical setting of our work: Airbnb platform and the regulation of short-term listings in the City of Los Angeles. Section 1.3 presents our dataset and the main variables of interest. Section 1.4 introduces our identification strategy. Section 1.5 illustrates our main results followed by a series of heterogeneity analysis, placebo tests and robustness checks. We conclude and discuss our results in Section 1.6.

6. The City of Los Angeles and the other 87 cities in the County are considered an interconnected housing market (Koster, Van Ommeren, and Volkhausen 2021; Bekkerman et al. 2023).

1.2 Background

1.2.1 Airbnb Platform

Airbnb is one of the leading digital platforms in the hospitality industry, connecting hosts with guests from all around the world for vocational rental. It was launched in 2008 and rapidly expanded, increasing the number of its users.

On the platform, guests can search for listings choosing over the number of days, period of the trip, and listings' characteristics (e.g., location, price). Then, a guest sends an inquiry for booking a stay for a given period to a host. Once the host accepts the request, the listing is booked for the selected period.

A feedback system enables guests and hosts to mutually review each other within 14 days after the end of the stay. A guest can leave one-to-five star ratings over different categories, write a public comment about the stay which will appear on the listing webpage, and write a private review to the host. Regarding the ratings, a guest can evaluate the overall experience, and rank separately other six subcategories: the accuracy of the listing description, the check-in process, the cleanliness of the listing, the communication with the host, the listing's location and the value-for-money. Only when a guest and a host have reviewed each other, or after 14 days, public comments are visible on the host's listing webpage on the platform and the rounded average of the scores is updated.[7]

1.2.2 Home-Sharing Ordinance in the City of Los Angeles

Many cities have adopted STR regulations in the last few years. The City of Los Angeles is not an exception. On December 2018, the city council approved the Home-Sharing Ordinance which was to be implemented and enforced in the next months:[8]

> "The Home-Sharing Ordinance will become effective on July 1, 2019. . . . Beginning July 1st ("implementation date"), hosts will be able to register for home-sharing using the City's online registration portal. Beginning November 1,

7. Further details about the Airbnb rating system can be found in Rossi 2023.

8. Refer to the text of HSO at https://clkrep.lacity.org/onlinedocs/2014/14-1635-S2_rpt_PLAN_06-13-2019.pdf

> 2019 ("enforcement date"), [...] the Department will begin overseeing enforcement of the ordinance [...]"

According to the HSO in the City of Los Angeles renting for periods shorter than 30 consecutive days is permitted only in the host's primary residence and for up to 120 days in a calendar year. Hosts are required to apply for a permit number and to communicate it to the Airbnb platform. Failing to provide the permit number, Airbnb announced the blocking of hosts from the platform.[9]

Figure 1 depicts the direct impact of the HSO. It plots the availability of residential properties listed for short-terms rent on Airbnb within the City of Los Angeles, as well as in other cities of the county.[10] During the implementation of the HSO, the number of short-term listings in the City of Los Angeles began to decline, and by the beginning of 2020, it had almost halved. Conversely, the other cities unaffected by similar housing policies exhibited a stable pattern, with no significant fluctuations in the supply of STR listings.

1.3 Data

1.3.1 *InsideAirbnb* Data

We collect data from *InsideAirbnb* for the county of Los Angeles. *InsideAirbnb* is a website that provides scraped data from Airbnb. It scrapes the platform on a regular basis, often once per month, and collects information about active listings. This includes fixed characteristics of the house (e.g., latitude and longitude of the listing) and time-varying characteristics (e.g., ratings, number of minimum nights for rent and prices). We match the latitude and longitude of each listing with city borders from Los Angeles city planning official statistics and define the location of each listing to clearly identify those within the City of Los Angeles.[11] We restrict the sample to listings that offer at minimum

9. Airbnb's announcement: https://www.airbnb.com/help/article/864/los-angeles-ca#shortterm

10. The county of Los Angeles includes the City of Los Angeles (green area in Figure 8 in Appendix 1.7) and other 87 cities and unincorporated areas (pink areas with black borders in Figure 8 in Appendix 1.7).

11. In Appendix 1.7, Figure 8 shows the location of Airbnb listings in the all county of Los Angeles. Most of the dwellings are located within the City of Los Angeles, followed by Santa Monica, Long Beach and West Hollywood.

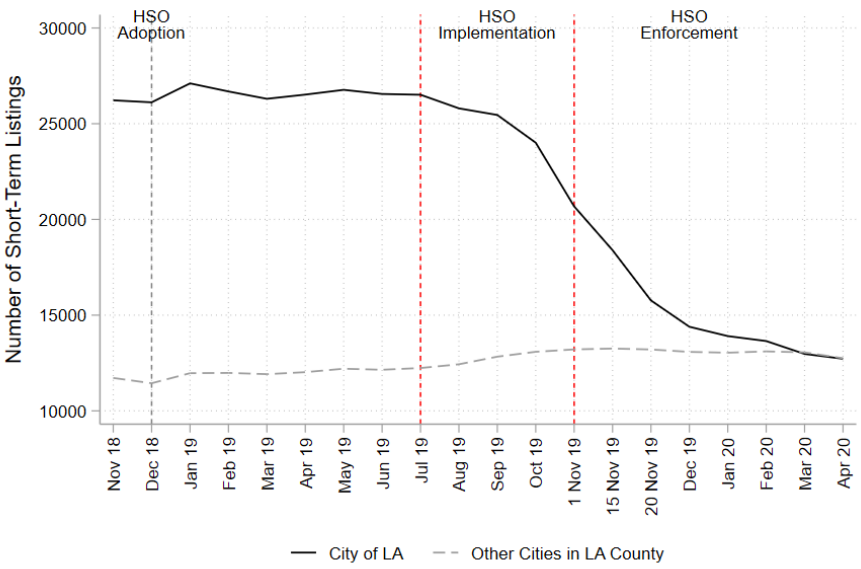


Figure 1: Number of short-term listings rented on Airbnb

Notes: The plot illustrates the evolution in the number of STR listings on Airbnb by scraping date (approximately each month). The solid line represents the number of STR listings rented on Airbnb within the City of Los Angeles. The line plots a remarkable decrease in the number of STR listings during the HSO implementation (1st of July - 1st of November 2019), and in the few first months after. The dashed line shows the more stable pattern of STR listings supply in the other 87 cities and unincorporated areas of the county of Los Angeles.

stays for less than 30 days. In this way, we can identify the short-term listings who were subject to the HSO rules. We additionally restrict the analysis to listings that exited the platform during the regulation implementation and entered before or at the very beginning of December 2018, before the policy was announced.[12] To check for parallel trends in the periods before the regulation, the estimation window is extended to fifteen snapshots before the implementation of the HSO.

1.3.2 Measurement of Host's Effort

We measure hosts' effort with their listing's ratings in related categories (i.e. check-in, communication and cleanliness ratings). Airbnb reports ratings on a five-point scale and displays the cumulative simple average over time. We observe $\bar{R}^k_{i,t}$ - the rounded cumulative average rating at time t for listing i for category k (e.g., communication,

12. We look at listings who left the platform between July and the beginning of November 2019. *InsideAirbnb* crawler scrapes platform data approximately once per month. In July 2019, scarping occurred on July 8th and 9th. At the beginning of November the platform was scraped on November 1st and 2nd.

check-in) - and $n_{i,t}$ - the cumulative number of new reviews of each listing i at time t. We compute the average rating per snapshot for each rating category - $r^k_{i,t}$ - as:

$$r^k_{i,t} = \frac{\bar{R}^k_{i,t} n_{i,t} - \bar{R}^k_{i,t-1} n_{i,t-1}}{(n_{i,t} - n_{i,t-1})} \tag{1}$$

It is important to notice that the accumulation of negative feedback and exit from the platform can influence each other. On the one hand, the benefit of exerting effort decreases when sellers are close to exit the platform and can anticipate their exit. These sellers may choose to "milk" their reputation and "shirk" in their last transactions. On the other hand, exit may not be anticipated, but actually driven by the accumulation of negative feedback in their last transactions on the platform. As a matter of fact, if sellers receive negative feedback, they may face challenges in attracting new guests and could be compelled to leave, as staying on the platform is no longer profitable. This reverse causality poses a challenge to the causal identification of the end-game considerations on sellers' effort. To overcome this issue, we exploit the implementation of the HSO as an exogenous shock in career concerns, to identify hosts who exited after a period of anticipation and for a policy that was in its design unrelated with hosts' reputation.

1.3.3 Analytical Sample

We start by identifying listings whose exit from the platform was independent from their hosts' accumulated reputation and could have been anticipated. In order to do so, we restrict the analysis to short-term listings whose last appearance in our data dates back to the HSO implementation period in the City of Los Angeles and that entered the platform before the regulation announcement in December 2018.[13] We assume that these listings exited due to the regulation either because ineligible or because the host was unwilling to pay to obtain the license number.

Table 1 presents summary statistics for our analytical sub-sample, made of 3,273 listings advertised by 2,209 hosts. The average number of listing's total reviews collected

13. We consider exit date the last date of appearance of a listing in our sample, meaning that the listing does not appear anymore on the platform anymore at least until December 2021.

over the lifetime of listings on the platform is 63. On a monthly basis, hosts accommodate approximately three guests (as approximated by the average listing's number of reviews per month) at an average price per night of $142 USD. Notably guests' ratings exhibit minimal variation, consistently maintaining a high average close to 5 stars across all categories. In half on the observations in our analytical sample, the number of hosts' listings advertised on the platform at the same time is above two. This observation suggests the presence of a significant proportion of "professional" hosts on the platform. These are individuals who rent out properties that are not their primary residences, deviating from the principle of the sharing economy. Finally, we complement platform data with administrative data from the 2018 Census, accounting for the share of ownership in households. We find that 50% of listings are located in areas with a limited share of owners (less than 20%).

Table 1: Distribution of variables in the analytical sample

	Mean	Median	S.D.	Min	Max	N.
Listing's tot. reviews ($n_{i,t}$)	63	39	69	2	706	28,773
Listing's n. reviews per month ($n_{i,t} - n_{i,t-1}$)	3.3	3	3	0	31	28,773
Listing's price per-night ($, USD)	142.1	115	118	10	1,599	28,773
Overall rating (#stars)	4.6	4.8	0.60	1	5	28,773
Accuracy rating (#stars)	4.7	5	0.68	1	5	28,773
Check-in rating (#stars)	4.8	5	0.59	1	5	28,773
Cleanliness rating (#stars)	4.6	5	0.77	1	5	28,773
Communication rating (#stars)	4.8	5	0.60	1	5	28,773
Location rating (#stars)	4.8	5	0.57	1	5	28,773
Value-for-money rating (#stars)	4.6	5	0.76	1	5	28,773
Host's listings (#number)	4.7	2	8	0	86	28,773
Owners in neighborhood (%)	26.2	19.5	23	0	94	28,752

Notes: The table summarizes key statistical moments (mean, median, standard deviation, minimum, maximum value and total number of data points) for the analyzed variables in a sample of 3,273 short-term listings in the City of Los Angeles. This sample includes listings that exited after the implementation of HSO, entered before its adoption, and have a price below $2,000 USD. For example, the table reports that the average total number of listing's reviews is 63. However, half of the observed listings have 39 or fewer reviews. The average value conceals significant variation, with a standard deviation of 69, which is larger than the mean value. The range of values spans from a minimum of 2 to a maximum of 706 reviews in a total number of 28,773 listing-snapshot observations.

1.4 Identification Strategy

On this selected analytical sample, we compare the evolution, before and after the HSO implementation, of ratings in effort-related categories with ratings on location. We select location as control for our Difference-in-Differences strategy, as it is a listing's attribute that is not affected by the policy as it is independent on hosts' effort. Figure 2 suggests indeed that, after controlling for individual characteristics, ratings on location did not significantly changes over the analyses period, while effort-related ratings declines after the regulation's announcement and adoption.

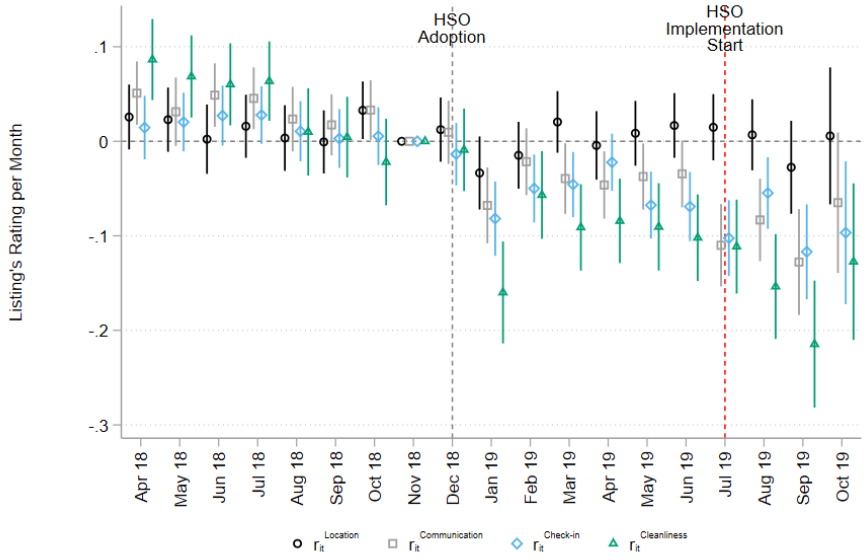


Figure 2: Evolution of ratings per listing overtime and by category

Notes: The figure plots the evolution of the residual (after controlling for listing fixed effect) average ratings in different categories compared to a reference period at the month before policy's implementation, together with the 95% confidence intervals. The sample is made of 3,273 short-term listings located in the City of Los Angeles that left the platform during the HSO implementation and entered before its adoption. The plot illustrates a significant drop in the average ratings for check-in (blue markers), cleanliness (green markers), and communication (gray markers) in the period after the policy adoption (December 2018) compared to the values in November 2018. Prior to the announcement, the average of effort-related ratings, especially for check-in and communication, remained relatively stable. Interestingly, the average rating for location (black markers) exhibits a stable pattern, showing no decline after the HSO announcement and during policy implementation.

To get causal estimates of the end-game effect, we estimate the following Equation:

$$r_{ikt} = \beta_1 after_t + \beta_2 CategoryEffort_{ikt} \times after_t + \delta_{it} + \mu_{ik} + \phi_{kt} + \epsilon_{ikt}, \qquad (2)$$

where r_{ikt} is the rating for listing i at snapshot t in a specific category k.

The indicator $CategoryEffort_{ikt}$ is equal to one if r_{ikt} measures effort, namely if it belongs to either the check-in, cleanliness or communication category and takes value zero if it relates to the listing's location.[14] The $after_t$ dummy is equal to one for all snapshots between June 2019 and the completion of the policy implementation and it is equal to zero in snapshots from April 2018 to May 2019. In the analysis, we gradually include a set of fixed effects to isolate the impact of confounders. We include a listing-month fixed effect δ_{it}, to harmonize for the same guest's feedback. Listing-category fixed effects are denoted by μ_{ik} and control for fixed characteristics of the listing in a specific category.[15] As it is indeed possible that seasonality affects guests' evaluation of location differently from the other categories, we control for the vector ϕ_{kt} which include category-month and category-year fixed effects. Standard errors are clustered at listing-category level to allow for correlation across snapshots for the same listing. Coefficient β_2 allows for the estimation of the effect of the end-game considerations on effort-related ratings.

Our identification strategy is based on the parallel trend assumption: in absence of the HSO, effort-related ratings and location ratings would have evolved in a parallel way. We adopt an Event Study framework to test this assumption by verifying the absence of significantly divergent pre-trends. We estimate:

$$r_{ikt} = \sum_{\tau=Apr18}^{Oct19} \beta_\tau CategoryEffort_{ikt} \times \mathbb{1}(t = \tau) + \delta_{it} + \mu_{ik} + \epsilon_{ikt} \tag{3}$$

1.5 Results

In this section, we report the estimates in Equation 2. Tables 2, 3 and 4 display the DiD parameters for the effort-related categories compared to location. In each column we progressively include the fixed effects.[16] Results point at an negative effect of end-game anticipation over effort. Ratings on check-in and communication decreased up to respectively -0.07 and -0.08 star points after the HSO implementation, equal to about

14. In each regression, each listing occurs twice per snapshot. For example, in the regression where we study the evolution of ratings for check-in, a listing i, at time t, occurs once for the value of r_{ikt} for check-in and once for the value of r_{ikt} for location.

15. μ_{ik} serves the scope of a more classical "individual fixed effect" in a standard DiD.

16. Coefficients tables for the other rating categories can be found in Tables 6 and 7 in Appendix 1.7.

1.5% of the average rating and of 12% of the standard deviation. The largest decline affected ratings on cleanliness, that declined by up to -0.1 points, equivalent to almost a 2.2% decline from the pre-announcement value of the rating and 13% of the standard deviation. The interpretation of our estimates' magnitude should take into account the sticky and J-shaped distribution of ratings on Airbnb.

Table 2: DiD estimates for the comparison of ratings on check-in with ratings on location

	(1) r_{ikt}	(2) r_{ikt}	(3) r_{ikt}	(4) r_{ikt}
Effort category (check-in)	0.039***	0.000	0.000	0.000
	(0.008)	(.)	(.)	(.)
After June 19	0.005	0.002	0.000	0.000
	(0.008)	(0.008)	(.)	(.)
Effort category (check-in) × After June 19	-0.073***	-0.066***	-0.066***	-0.060***
	(0.013)	(0.013)	(0.012)	(0.021)
Listing-Category FE		✓	✓	✓
Listing-month FE			✓	✓
Month-Category FE				✓
Year-Category FE				✓
Standard Errors Clustering Level	listing-category	listing-category	listing-category	listing-category
R^2	0.002	0.271	0.688	0.689
Number of observations	57,546	57,546	57,546	57,546

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients estimates from Equation 2 when the effort-related category under study (compared to location) is check-in. Standard errors are clustered at listing-category level and reported in parentheses. In column (2) we include listing-category fixed effect, in column (3) we add the listing-month fixed effect and in column (4) also the year-category fixed effect. In all the specifications the coefficient for the interaction of effort category and after June 2019 is negative and significant at 1% level. The magnitude of the end-game effect (coefficient of the interaction) range between -0.060 to -0.073 star points.

The event study, illustrated by Figure 3 supports our identification strategy and shows no significant pre-trend in the difference between ratings on check-in and ratings on location before the HSO approval and its consequent announcement in the media after December 2018.[17] The point estimates of coefficients β_τ from Equation 3 start to become significantly negative in periods after the HSO approval, and stayed below zero during its implementation.[18]

17. Example of the HSO approval coverage in the media: https://www.latimes.com/local/lanow/la-me-ln-airbnb-rental-ordinance-20181211-story.html

18. ES estimates for the other rating effort-related ratings categories are reported in Figures 9 and 10 in Appendix 1.7. In addition, Figures 11 and 12 report β_τ from Equation 3 also for ratings on accuracy and value-for-money.

Table 3: DiD estimates for the comparison of ratings on communication with ratings on location

	(1) r_{ikt}	(2) r_{ikt}	(3) r_{ikt}	(4) r_{ikt}
Effort category (communication)	0.042***	0.000	0.000	0.000
	(0.008)	(.)	(.)	(.)
After June 19	0.005	0.002	0.000	0.000
	(0.008)	(0.008)	(.)	(.)
Effort category (communication) × After June 19	-0.079***	-0.078***	-0.078***	-0.056***
	(0.013)	(0.013)	(0.012)	(0.021)
Listing-Category FE		✓	✓	✓
Listing-month FE			✓	✓
Month-Category FE				✓
Year-Category FE				✓
Standard Errors Clustering Level	listing-category	listing-category	listing-category	listing-category
R^2	0.002	0.264	0.681	0.682
Number of observations	57,546	57,546	57,546	57,546

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients estimates from Equation 2 when the effort-related category under study (compared to location) is communication. Standard errors are clustered at listing-category level and reported in parentheses. In column (2) we include listing-category fixed effect, in column (3) we add the listing-month fixed effect and in column (4) also the year-category fixed effect. In all the specifications the coefficient for the interaction of effort category and after June 2019 is negative and significant at 1% level. The magnitude of the end-game effect (coefficient of the interaction) range between -0.056 to -0.079 star points.

Table 4: DiD estimates for the comparison of ratings on cleanliness with ratings on location

	(1) r_{ikt}	(2) r_{ikt}	(3) r_{ikt}	(4) r_{ikt}
Effort category (cleanliness)	-0.171***	0.000	0.000	0.000
	(0.010)	(.)	(.)	(.)
After June 19	0.005	0.002	0.000	0.000
	(0.008)	(0.008)	(.)	(.)
Effort category (cleanliness) × After June 19	-0.092***	-0.104***	-0.104***	-0.026
	(0.014)	(0.014)	(0.013)	(0.025)
Listing-Category FE		✓	✓	✓
Listing-month FE			✓	✓
Month-Category FE				✓
Year-Category FE				✓
Standard Errors Clustering Level	listing-category	listing-category	listing-category	listing-category
R^2	0.021	0.270	0.669	0.671
Number of observations	57,546	57,546	57,546	57,546

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients estimates from Equation 2 when the effort-related category under study (compared to location) is cleanliness. Standard errors are clustered at listing-category level and reported in parentheses. In column (2) we include listing-category fixed effect, in column (3) we add the listing-month fixed effect and in column (4) also the year-category fixed effect. In columns (1), (2) and (3) the coefficient for the interaction of effort category and after June 2019 is negative and significant at 1% level. Including the year fixed effect cancels the significance of the effect; however, the coefficient still remains negative. The magnitude of the end-game effect (coefficient of the interaction) range between -0.026 to -0.104 star points.

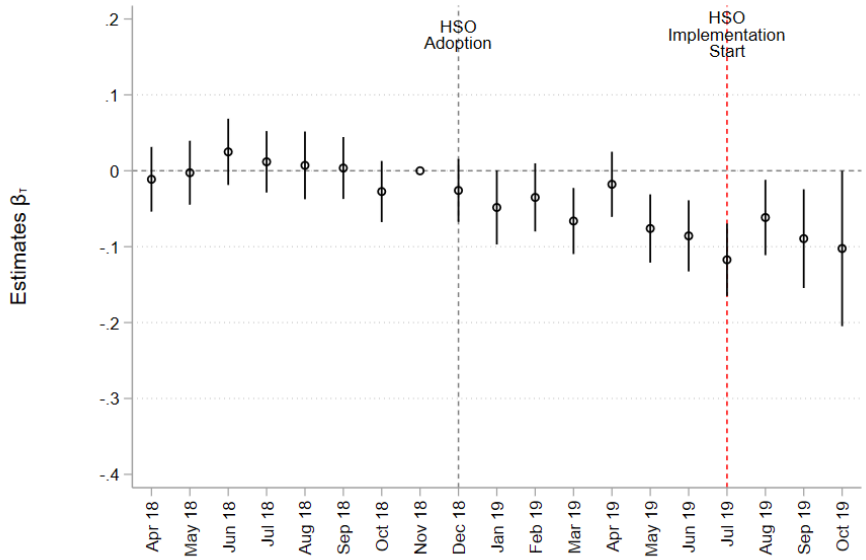


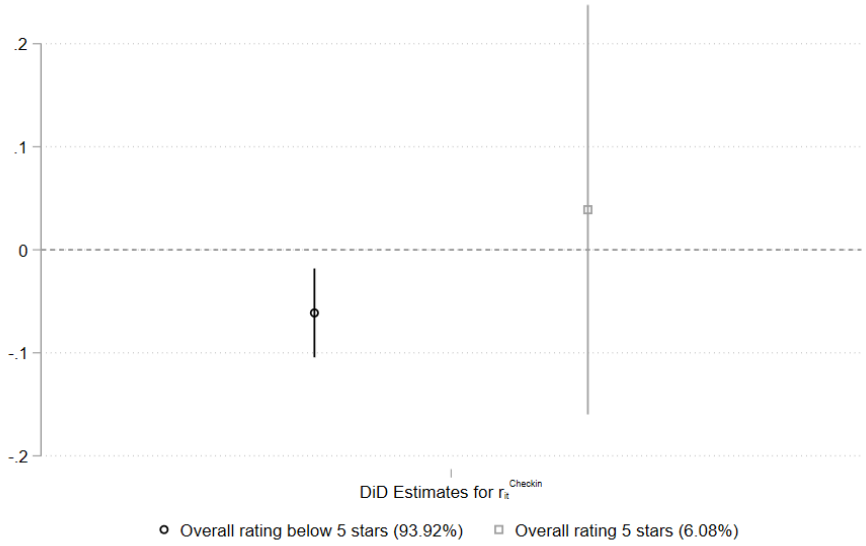
Figure 3: Event Study estimates comparing ratings on check-in with ratings on location

Notes: The figure plots the estimates for coefficients β_τ in Equation 3 along with the 95% confidence intervals for each estimate. Number of observations is 57,546 (each listing is observed twice: once for the effort-related rating, once for location). The reference period, corresponding to the month before HSO approval, is normalized to zero. Standard errors are clustered at a month-listing level. It is important to note that standard errors increased in the last snapshots of data due to the loss of observations when some listings left the platform after July 2019.

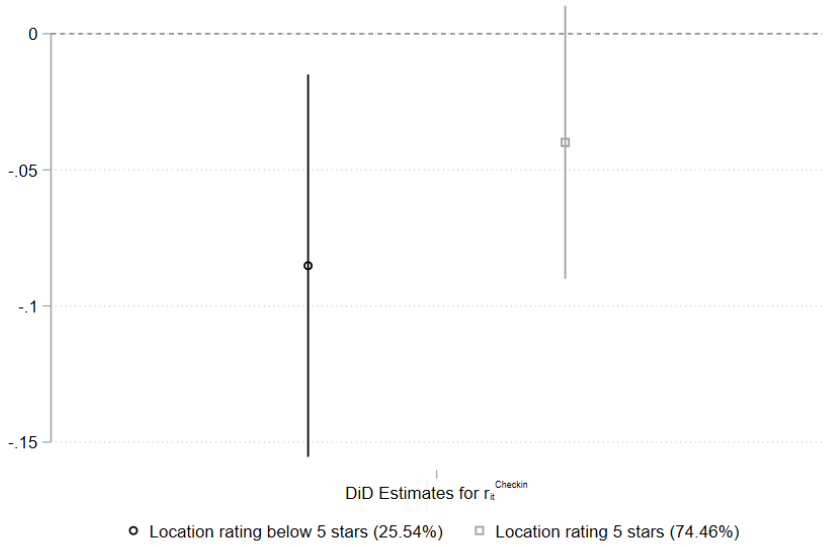
1.5.1 Heterogeneity Analysis

In this subsection, we present the results of a heterogeneity analysis aimed at examining the moderation effect of various hosts' and listings' characteristics on the impact of end-game considerations over effort decisions. We begin by investigating the moderation influence of hosts' previous reputation. Even if hosts can anticipate their forthcoming exit, they still face a period of time before they are forced to leave the platform when they can make profit by attracting guests. Hosts with higher overall ratings or those advertising listings in central or more touristic areas may continue be attractive to guests, even if their ratings start to decline and therefore have larger margins to leverage accumulated ratings for continuing their business. However, as depicted in Figure 4, there is no significant difference in the estimates of β_2 (from Equation 2) given the cumulative average rating of listings before the HSO approval. This holds true for both the overall rating (a) and the location-specific one (b).[19]

19. Figures 13 and 14 in Appendix 1.7 show the heterogeneity of the end-game effect by the value of the overall rating before HSO, respectively on communication and cleanliness rating. Figures 15 and 16



(a) Heterogeneity analysis by listing's overall rating before the HSO approval



(b) Heterogeneity analysis by listing's rating on location before the HSO approval

Figure 4: Heterogeneity analysis on the check-in ratings by listing's reputation before HSO approval

Notes: The figures plot the estimates of β_2 from Equation 2 for ratings on check-in and the 95% confidence interval for the estimates. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given (a) the value of $\bar{R}_{it}^{overall}$ and (b) the value of $\bar{R}_{it}^{location}$ before December 2018. Although the decrease in effort is more pronounced for listings with lower reputation before the policy announcement, these differences are not statistically significant.

Finally, to strengthen the validity of our results, we focus on characteristics related

in Appendix 1.7 illustrate the heterogeneity of the end-game effect by the value of the location rating before HSO, respectively on communication and cleanliness rating.

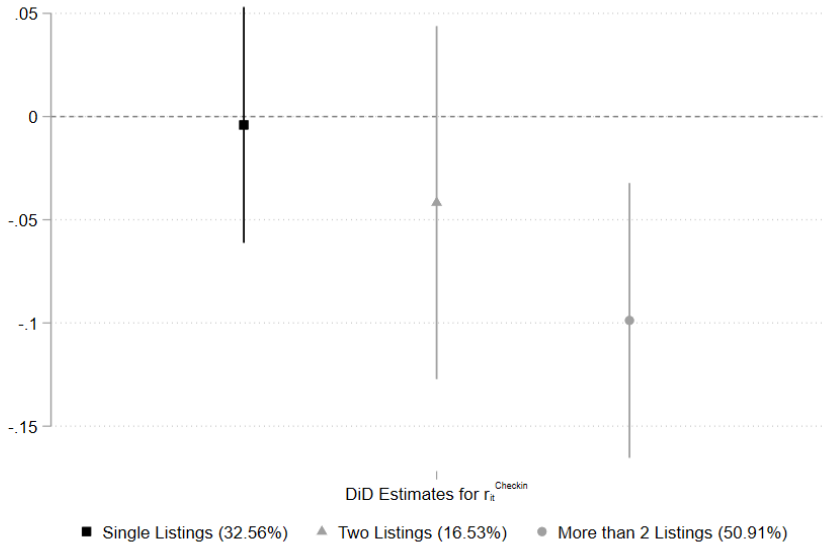
to the regulatory requirements, specifically examining the situation of those hosts who exited likely because their listings were deemed ineligible. The regulation stipulated that only primary residences could be rented for STR purposes. This implies that the target seller consisted mostly of "professional" hosts. We approximate the likelihood of being ineligible in two ways. First, we identify professional hosts by the number of listings rented out on Airbnb prior to the HSO implementation. We assume that having more than one listing would increase the likelihood of being professional, meaning renting for short-term a dwelling which is not the host's primary residence.[20] Then, we look at the share of home owners in the neighborhood where the listing is located. If the share is very low and there is a higher prevalence of dwellings being rented out, it is likely that the listing is not the primary residence of the owner. Our findings at Figure 5 indicate indeed that listings rented by with multiple-listing hosts (a) and those located in areas with fewer property owners (b), exhibited a more pronounced decline in effort-related ratings during the HSO implementation. These results highlight a greater focus on profit maximization among "professional". Conversely, the intrinsic value associated with participating in the sharing economy for non-professional individuals could partially mitigate the moral hazard at the end of hosts' career.[21]

1.5.2 Robustness Checks and Placebos

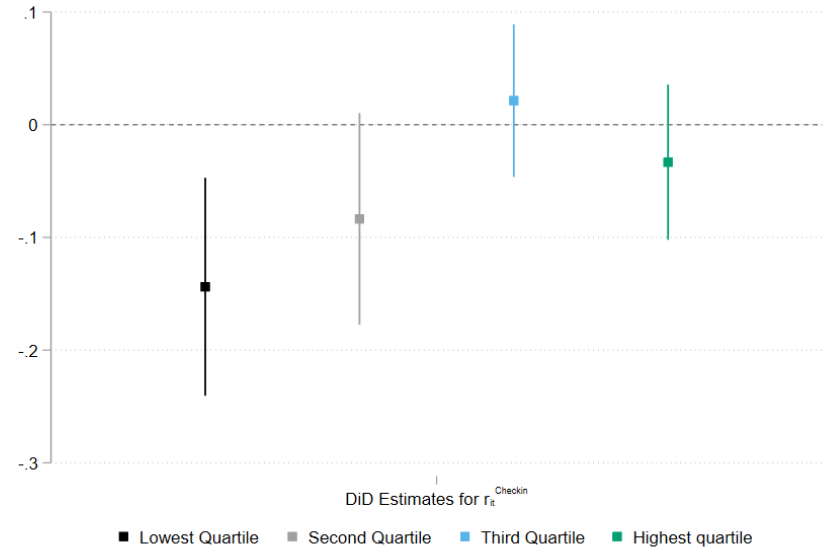
In addition to our main results, we perform a set of robustness checks and placebo analysis. First, we show that our results are robust to different clustering of the standard errors and dataset restrictions (Table 8 in Appendix 1.7). Additionally, we look at the evolution of other measures of effort such as the host's response rate to guest messages and requests and the times it takes for them to answer. The descriptive results presented in Figure 6 resonates with the main findings of an effort decline. In last hosts' transactions,

20. It is however possible that two listings of the same host are located in the same house or apartment. In this study we do not distinguish between the two possibilities.

21. Figures 17 and 18 in Appendix 1.7 show the heterogeneity of the end-game effect by the host's number of listings on Airbnb in Los Angeles before the HSO approval, respectively on communication and cleanliness rating. Figures 19 and 20 in Appendix 1.7 illustrate show the heterogeneity of the end-game effect by the share of owners in the listing's neighbourhood, respectively on communication and cleanliness rating.



(a) Heterogeneity analysis by host's number of listings on Airbnb before the HSO approval

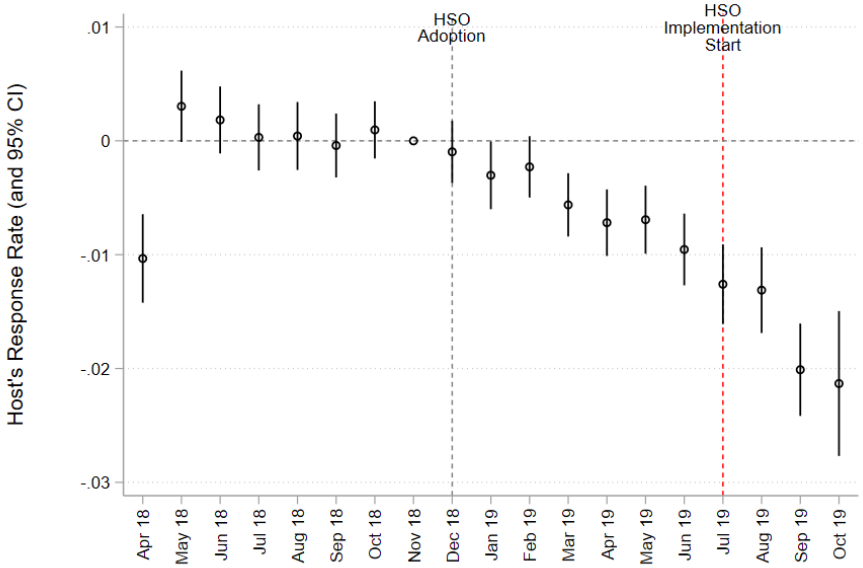


(b) Heterogeneity analysis by share of owners in the listing's neighbourhood

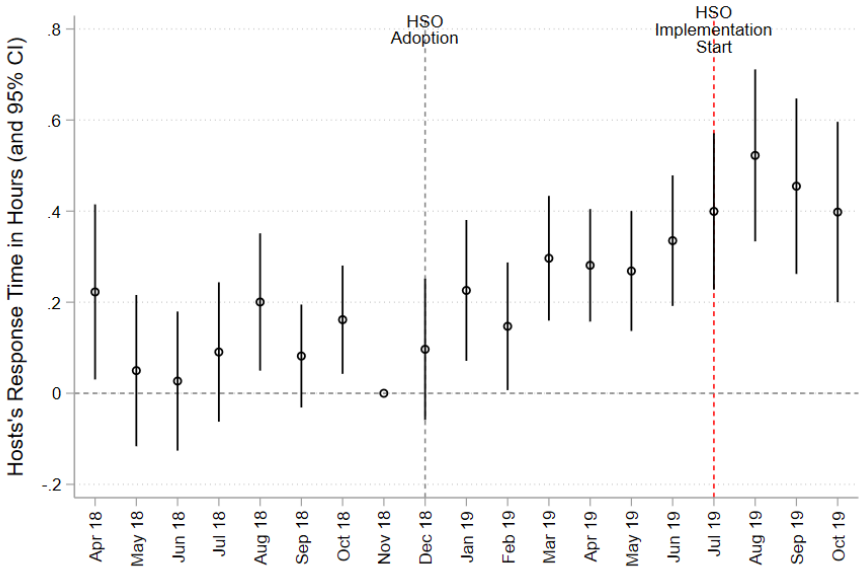
Figure 5: Heterogeneity analysis on the rating on check-in by eligibility conditions

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on check-in. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given (a) the host's number of listings (in the county of Los Angeles) on the platform before the HSO approval and (b) the share of owners in listing's neighbourhood according to the 2018 census. Sub-Figure (a) illustrates that, while not particularly statistically significant, the end-game effect is more prominent among listings hosted by individuals with more than 2 rooms or houses listed on Airbnb before HSO approval. In sub-Figure (b), it is observed that listings located in neighborhoods with a low share of owners (lower quartile) exhibit a more pronounced decline in effort compared to the other three quartiles. The difference is significant in particular in comparison with the upper half of the owners share distribution.

the response rate decreased among sellers who left the platform in City of Los Angeles due to the regulation, while their time to answer requests and messages increased.



(a) Evolution of hosts' response rate



(b) Evolution of hosts' response time

Figure 6: Evolution of alternative measures of hosts' effort

Notes: The figure plots the evolution of residual average value of host's response rate (a) and response time (measured in hours), measure in hours (b) compared to reference period (the month before policy implementation). Vertical bars represent the the 95% confidence interval. The sample is made of 3,273 short-term listings located in the City of Los Angeles that left the platform during the HSO implementation and entered before its adoption. Sub-Figure (a) demonstrates a decline in the average hosts' response rate from the moment the policy is announced compared to November 2018. Conversely, Sub-Figure (b) reveals a statistically significant increase in the average hosts' response time to guest messages after the HSO adoption.

Finally, in Figure 7, we perform a placebo test looking at listings located in other cities of the county, thus not affected by the HSO but which showed similar characteristics and left the platform in the same period. As expected, we find no significant changes in the difference between the effort-related ratings and ratings on location.

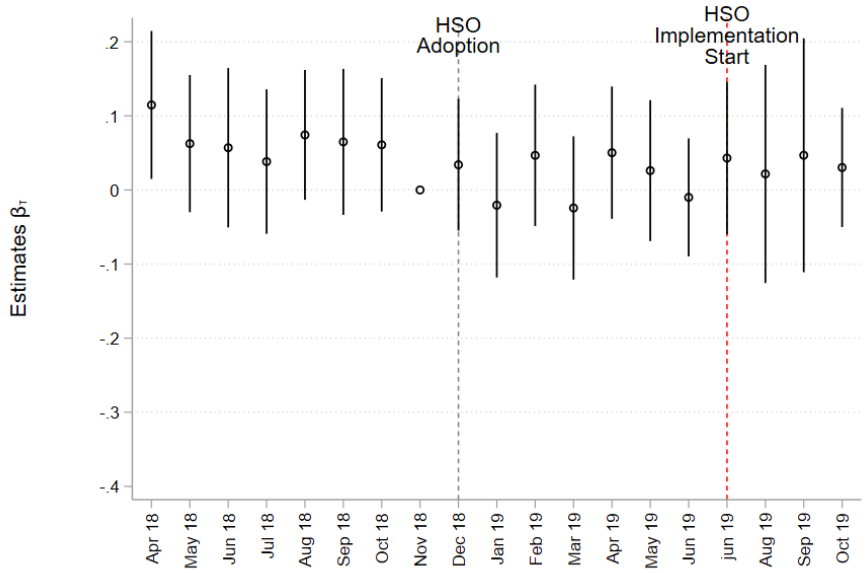


Figure 7: Placebo Event study estimates comparing ratings on check-in with ratings on location in other cities in the county of Los Angeles

Notes: The figure plots the estimates for coefficients β_τ in Equation 3 for listings outside of the City of Los Angeles but belonging to the same county and having similar features (STR listings exited during the implementation of HSO, entered before its adoption, and have a price below $2,000 USD). We show no significant changes in the DiD coefficient after the HSO is announced and during its implementation.

1.5.3 Additional Specification

We finally complement our approach with a more traditional DiD. We take advantage of the geographical variation in the HSO requirement and the richness of our data which covers listings in the whole county of Los Angeles including cities outside the administrative borders of the City of Los Angeles. The targeted sample includes listings who exited the platform during the policy implementation. The treated group corresponds with listings located in the City of Los Angeles. The control group includes all the remaining cities. This division is based on the idea that hosts who are not eligible for registration, or unwilling to register, are more likely to anticipate their unavoidable exit after regulation is enforced. Differently, hosts who are not subject to regulation can still

exit the platform but are unlikely to anticipate it so much in advance. Moreover, we focus on listings leaving the platform simultaneously to rule out the risk of spillovers due to decreased competition driven by sudden exit of many listings on the platform.[22]

The main equation to estimate is:

$$r_{it}^{k} = \beta_1 after_t + \beta_2 LAcity_i \times after_t + \phi X_{it} + \mu_i + \delta_t' + \gamma Trend_t + \epsilon_{it}, \tag{4}$$

where r_{it}^{k} is the rating for listing i at snapshot t and category k. The indicator $LAcity_i$ identified the treated group, namely listings whose exit could have been anticipated by host. It is equal to one if listing i located in the City of Los Angeles, and it is equal to zero if a listing is located in other cities of the county. As in the main specification, $after_t$ takes value 1 after June 2019. In the analysis, we gradually include controls to isolate the impact of confounders. To account for seasonality, we control for the vector, δ_t', which includes the set of month and year dummies, and for $Trend_t$, which is a daily linear trend. X_{it} includes the total number of reviews received by each listing since entry on the platform. Listing fixed effects are denoted by μ_i. Standard errors are clustered at a listing level to allow for correlation across snapshots for the same listing.

We assume that, in absence of HSO, the evolution of ratings of exiters among listings located in the City of Los Angeles and those outside the city's administrative border would have been the same. We adopt an event study approach and show the plausibility of this assumption in our context (Figures 21, 22 and 23 in Appendix 1.7).

In Table 5, we document a statistically significant negative coefficient for both check-in and communication ratings. Including all the controls, anticipation decreases a rating by -0.04 points for check-in and -0.06 for communication. As for the main specification results, coefficient magnitude should be interpreted given the limited variance of ratings on the platform. The absence of a significant effect on cleanliness ratings may be interpreted as being somehow related to the possibility that these services are often outsourced to external cleaning companies. Possibly a loss of clients of these companies

22. Table 9 in Appendix 1.7 summarizes the distribution of the variables used in the analysis, comparing their mean in the treated and the control group.

Table 5: DiD estimates of β_2 from Equation 4

	(1) $r^{Check-in}$	(2) $r^{Cleanliness}$	(3) $r^{Communication}$	(4) $r^{Overall}$	(5) $r^{Accuracy}$	(6) $r^{Location}$	(7) r^{Value}	(8) Price
After June 19	0.010	0.042	0.058**	0.001	0.031	0.038	-0.021	0.875
	(0.026)	(0.034)	(0.023)	(0.027)	(0.033)	(0.026)	(0.036)	(1.368)
City of LA × After June 19	-0.043**	-0.021	-0.062***	-0.014	-0.039	0.005	-0.004	-1.808
	(0.022)	(0.029)	(0.019)	(0.023)	(0.028)	(0.021)	(0.031)	(1.411)
Listing-Month Controls (X'_{it})	✓	✓	✓	✓	✓	✓	✓	✓
Time Linear Trend	✓	✓	✓	✓	✓	✓	✓	✓
Time FEs	✓	✓	✓	✓	✓	✓	✓	✓
Listing FE	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.278	0.264	0.275	0.273	0.258	0.286	0.217	0.978
Number of observations	34,279	34,279	34,279	34,279	34,279	34,279	34,279	34,279

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients β_2 from Equation 4. Standard errors are clustered listing level and are reported in parentheses.

due to exit of hosts affected by the HSO, impacted their business also in other cities in the area. Meanwhile, the lack of an effect on location ratings confirms the validity of our choice for the control group in the main specification. Finally, while the anticipation of a forthcoming career termination correlates with lower effort exertion, we do not find evidences of moral hazard passing through changes in pricing strategy.

1.6 Conclusions

In this work, we provide empirical evidence of moral hazard when sellers can anticipate the end of their careers on a digital platform. Using a panel of listings present on Airbnb in City of Los Angeles, we study hosts' effort decisions in response to a regulatory shock. The policy generated anticipation of exit among a group of hosts advertising short-term-rentals on the platform. The regulation introduced additional costs and eligibility criteria for hosts to rent form short-term periods, leading to a significant drop in the number of listings present on Airbnb during the policy implementation, from July to November 2019. When the policy was announced, hosts could anticipate if they were going or not to abandon the platform in the next months. Accordingly, we focus on listings affected by the regulation that left the platform during its implementation and compare the evolution of ratings reflecting hosts' effort with ratings on location. We also apply an additional Difference-in-Differences strategy and compare listings located in the City of Los Angeles

with listings located in other cities of the county, where the regulation did not apply.

With both identification strategies, we document a negative and significant impact of end-game considerations on ratings reflecting hosts' effort. When hosts expect few remaining periods on the platform, they tend to shirk and in turn, their effort-related ratings decrease. Our results suggest that reputation incentives vary across sellers' life on the platform and specifically, they have less incentives to exert more effort when they approach the end of their careers. The specific context of Airbnb reveals that reputation systems relying solely on the simple averaging of ratings over the entire sellers' lifespan are inadequate in mitigating moral hazard during sellers' final transactions on digital platforms. Such evidences pave the way to important managerial discussion around reputation systems design. Platforms could mitigate the risk moral hazard in the last transactions of sellers, by introducing additional tools (e.g., performance-pay contracts, higher weights to more recent feedback, etc.). They also seem to suggest that measures such as censoring ratings after a certain number of reviews may be detrimental and stimulate hidden actions by sellers, given the decreasing marginal benefits of a positive rating after.

1.7 Appendix

Additional Results

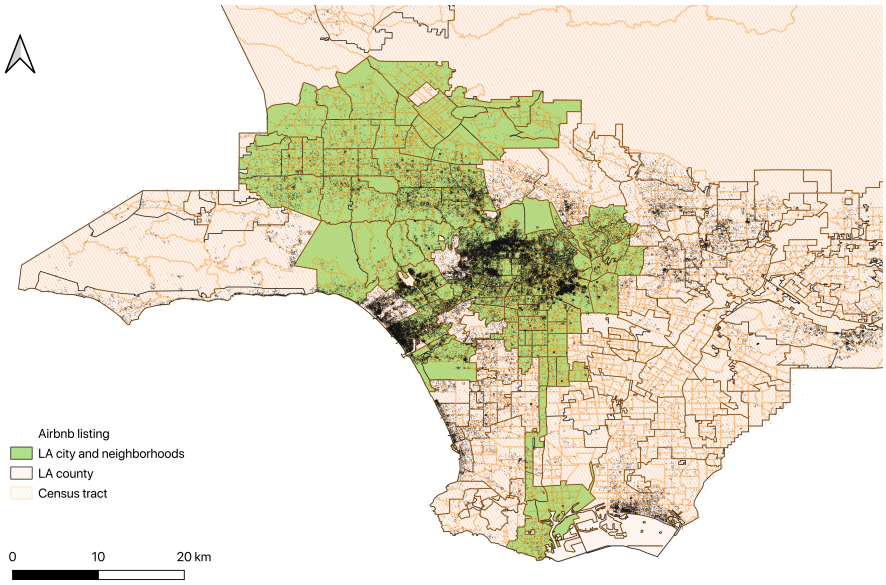


Figure 8: Location of Airbnb listings in the county of Los Angeles

Notes: Authors' own computations based on *InsideAirbnb* data. The figure displays the locations of all listings present in the *InsideAirbnb* dataset. The green area represents the city of Los Angeles, while the pink area between borders encompasses other regions within the county of Los Angeles. The majority of dwellings are situated within the City of Los Angeles, followed by Santa Monica, Long Beach, and West Hollywood.

Table 6: DiD estimates for the comparison of ratings on accuracy with ratings on location

	(1) r_{ikt}	(2) r_{ikt}	(3) r_{ikt}	(4) r_{ikt}
Category (accuracy)	-0.031***	0.000	0.000	0.000
	(0.009)	(.)	(.)	(.)
After June 19	0.005	0.002	0.000	0.000
	(0.008)	(0.008)	(.)	(.)
Category (accuracy) × After June 19	-0.099***	-0.102***	-0.102***	-0.060***
	(0.014)	(0.014)	(0.013)	(0.022)
Listing-Category FE		✓	✓	✓
Listing-month FE			✓	✓
Month-Category FE				✓
Year-Category FE				✓
Standard Errors Clustering Level	listing-category	listing-category	listing-category	listing-category
R^2	0.004	0.257	0.678	0.678
Number of observations	57,546	57,546	57,546	57,546

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients estimates from Equation 2 when the effort-related category under study (compared to location) is accuracy. Standard errors are clustered at listing-category level and reported in parentheses. In column (2) we include listing-category fixed effect, in column (3) we add the listing-month fixed effect and in column (4) also the year-category fixed effect. In all the specifications the coefficient for the interaction of effort category and after June 2019 is negative and significant at 1% level. The magnitude of the end-game effect (coefficient of the interaction) range between -0.060 to -0.102 star points. Although less directly associated with effort, accuracy ratings also decline after the HSO approval, while there were no significant differences before it. This suggests that accuracy ratings can be interpreted as a measure of host effort. For instance, hosts anticipating an exit may not be spending time updating descriptions of the listing that may include outdated information about appliance and dwelling status or other details.

Table 7: DiD estimates for the comparison of ratings on value-for-money with ratings on location

	(1) r_{ikt}	(2) r_{ikt}	(3) r_{ikt}	(4) r_{ikt}
Category (value-for-money)	-0.181***	0.000	0.000	0.000
	(0.009)	(.)	(.)	(.)
After June 19	0.005	0.002	0.000	0.000
	(0.008)	(0.008)	(.)	(.)
Category (value-for-money) × After June 19	-0.117***	-0.121***	-0.121***	-0.057**
	(0.015)	(0.015)	(0.013)	(0.025)
Listing-Category FE		✓	✓	✓
Listing-month FE			✓	✓
Month-Category FE				✓
Year-Category FE				✓
Standard Errors Clustering Level	listing-category	listing-category	listing-category	listing-category
R^2	0.025	0.242	0.670	0.671
Number of observations	57,546	57,546	57,546	57,546

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients estimates from Equation 2 when the effort-related category under study (compared to location) is value-for-money. Standard errors are clustered at listing-category level and reported in parentheses. In column (2) we include listing-category fixed effect, in column (3) we add the listing-month fixed effect and in column (4) also the year-category fixed effect. In the first three specifications the coefficient for the interaction of effort category and after June 2019 is negative and significant at 1% level, significance decreases to 5% in column (4). The magnitude of the end-game effect (coefficient of the interaction) range between -0.057 to -0.121 star points. Since Table 3 shows that end-game concerns do not affect price, the decline in value-for-money can be interpreted as a negative evaluation over sellers' effort.

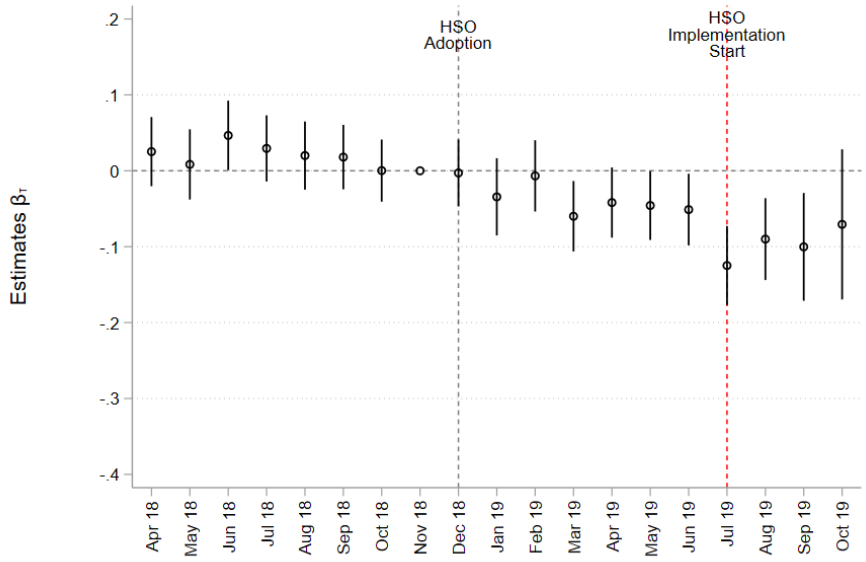


Figure 9: Event Study estimates comparing ratings on communication with ratings on location

Notes: The figure plots the estimates for coefficients β_τ in Equation 3 along with the 95% confidence intervals for each estimate. Number of observations is 57,546 (each listing is observed twice: once for the effort-related rating, once for location). The reference period, corresponding to the month before HSO approval, is normalized to zero. Standard errors are clustered at a month-listing level. It is important to note that standard errors increased in the last snapshots of data due to the loss of observations when some listings left the platform after July 2019.

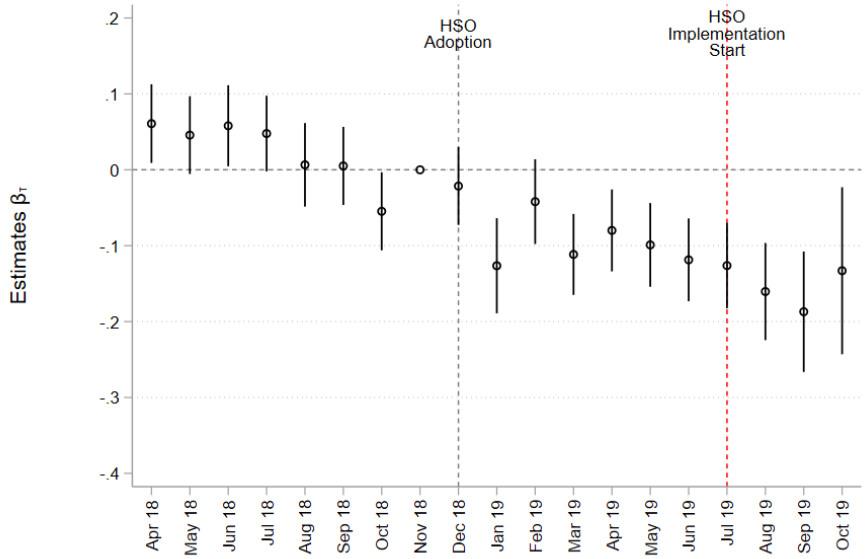


Figure 10: Event Study estimates comparing ratings on cleanliness with ratings on location

Notes: The figure plots the estimates for coefficients β_τ in Equation 3 along with the 95% confidence intervals for each estimate. Number of observations is 57,546 (each listing is observed twice: once for the effort-related rating, once for location). The reference period, corresponding to the month before HSO approval, is normalized to zero. Standard errors are clustered at a month-listing level. It is important to note that standard errors increased in the last snapshots of data due to the loss of observations when some listings left the platform after July 2019.

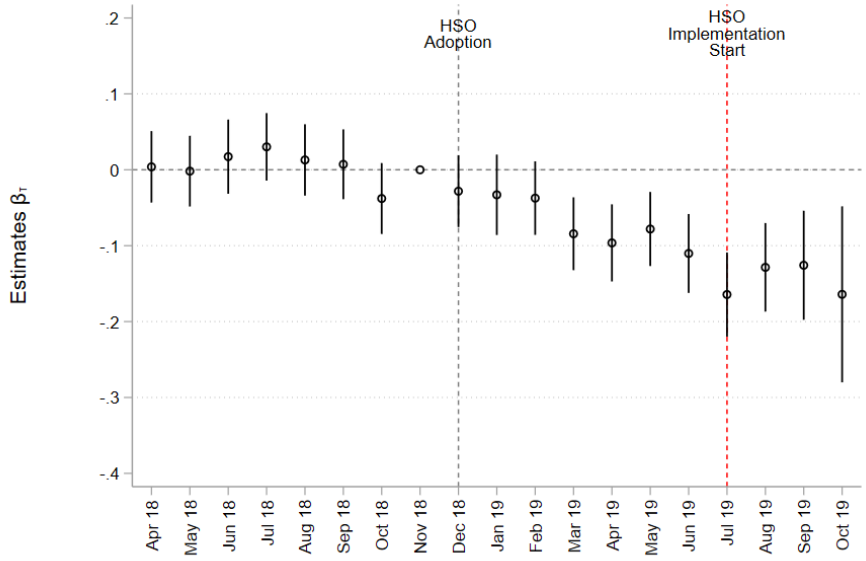


Figure 11: Event Study estimates comparing ratings on accuracy with ratings on location

Notes: The figure plots the estimates for coefficients β_τ in Equation 3 along with the 95% confidence intervals for each estimate. Number of observations is 57,546 (each listing is observed twice: once for the effort-related rating, once for location). The reference period, corresponding to the month before HSO approval, is normalized to zero. Standard errors are clustered at a month-listing level. It is important to note that standard errors increased in the last snapshots of data due to the loss of observations when some listings left the platform after July 2019.

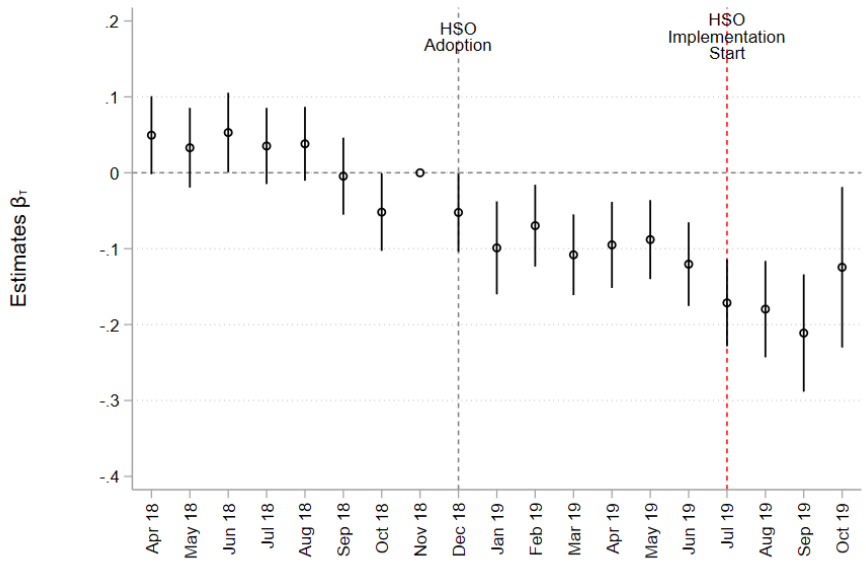


Figure 12: Event Study estimates comparing ratings on value-for-money with ratings on location

Notes: The figure plots the estimates for coefficients β_τ in Equation 3 along with the 95% confidence intervals for each estimate. Number of observations is 57,546 (each listing is observed twice: once for the effort-related rating, once for location). The reference period, corresponding to the month before HSO approval, is normalized to zero. Standard errors are clustered at a month-listing level. It is important to note that standard errors increased in the last snapshots of data due to the loss of observations when some listings left the platform after July 2019.

Additional Robustness and Heterogeneity Analysis

Table 8: DiD estimates robustness to different specifications and sampling

	(1) r_{ikt}	(2) r_{ikt}	(3) r_{ikt}	(4) r_{ikt}	(5) r_{ikt}	(6) r_{ikt}
Effort category (check-in)	0.000	0.000	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)	(.)	(.)
After June 19	0.000	0.000	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)	(.)	(.)
Effort category (check-in) × After June 19	-0.060***	-0.056***	-0.060***	-0.060***	-0.059***	-0.055**
	(0.021)	(0.018)	(0.021)	(0.021)	(0.021)	(0.023)
Listing-Category FE	✓	✓	✓	✓	✓	✓
Listing-Time FE	✓	✓	✓	✓	✓	✓
Month-Category FE	✓	✓	✓	✓	✓	✓
Year-Category	✓	✓	✓	✓	✓	✓
Sample				price<500	price<1000	distance≥0.003
Standard Errors Clustering Level	listing-category	zip-code	ct10	listing-category	listing-category	listing-category
R^2	0.689	0.689	0.689	0.687	0.688	0.688
Number of observations	57,546	57,042	57,546	56,662	57,426	49,524

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table tests the robustness of the coefficients estimated from Equation 5 (column (1)) to different clustering levels of the standard errors (column (2) and column (3) and to different samples (column (4), column (5) and column (6)). Standard errors are reported in parentheses.

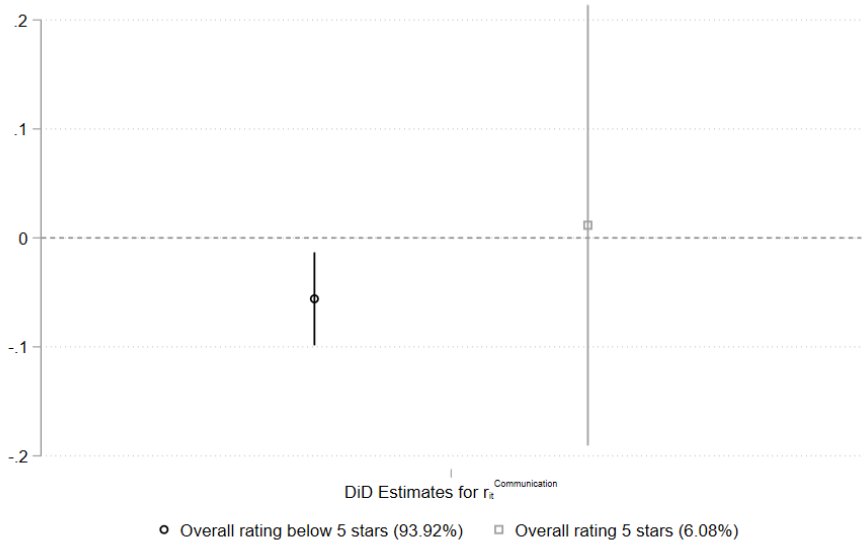


Figure 13: Heterogeneity analysis on ratings on communication by listing's overall rating before the HSO approval

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on communication and the estimates' 95% confidence intervals. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the value of $\bar{R}_{it}^{overall}$ before December 2018.

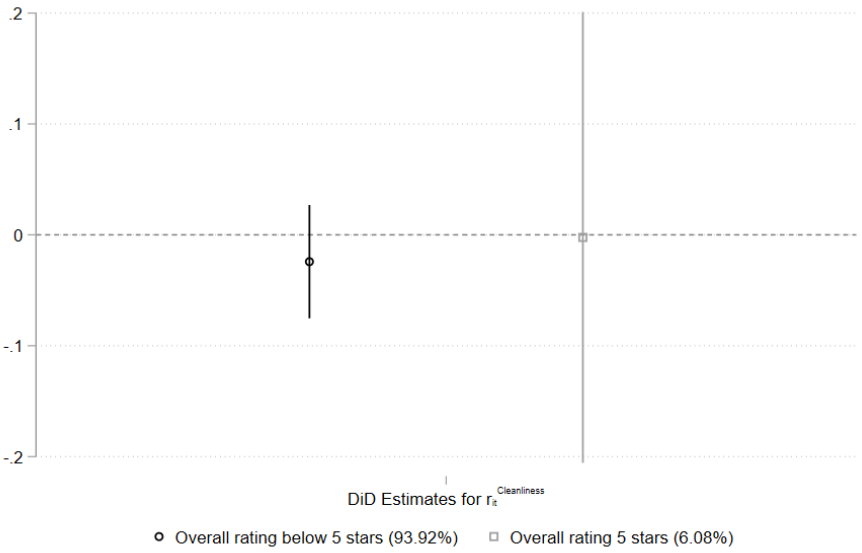


Figure 14: Heterogeneity analysis on ratings on cleanliness by listing's overall rating before the HSO approval

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on cleanliness. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the value of $\bar{R}_{it}^{overall}$ before December 2018.

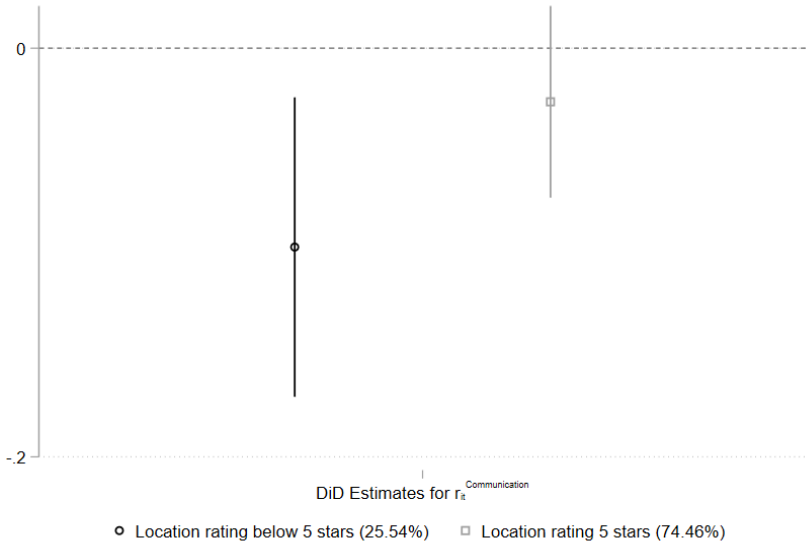


Figure 15: Heterogeneity analysis on ratings on communication by listing's rating on location before the HSO approval

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on communication and the estimates' 95% confidence intervals. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the value of $\bar{R}_{it}^{location}$ before December 2018.

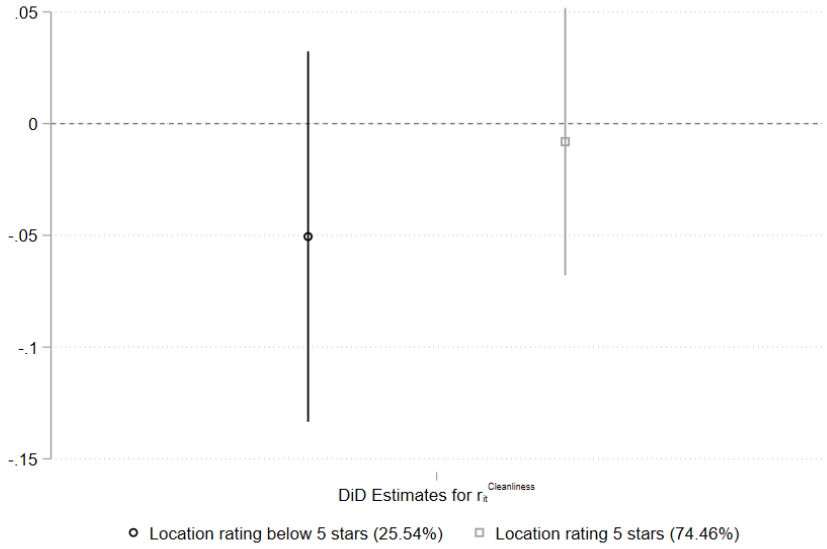


Figure 16: Heterogeneity analysis on ratings on cleanliness by listing's rating on location before the HSO approval

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on cleanliness and the estimates' 95% confidence intervals. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the value of $\bar{R}_{it}^{location}$ before December 2018.

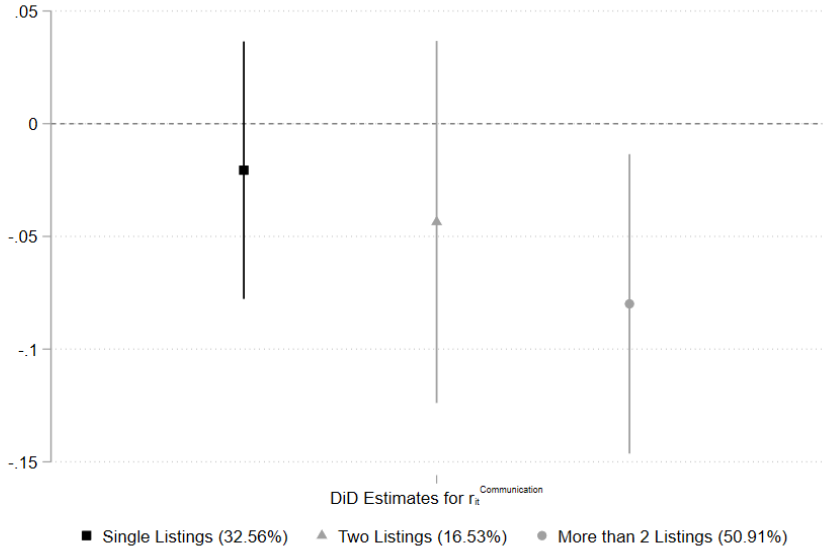


Figure 17: Heterogeneity analysis on ratings on communication by host's number of listings on Airbnb before the HSO approval

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on communication and the estimates' 95% confidence intervals. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the host's number of listings on the platform before the HSO approval.

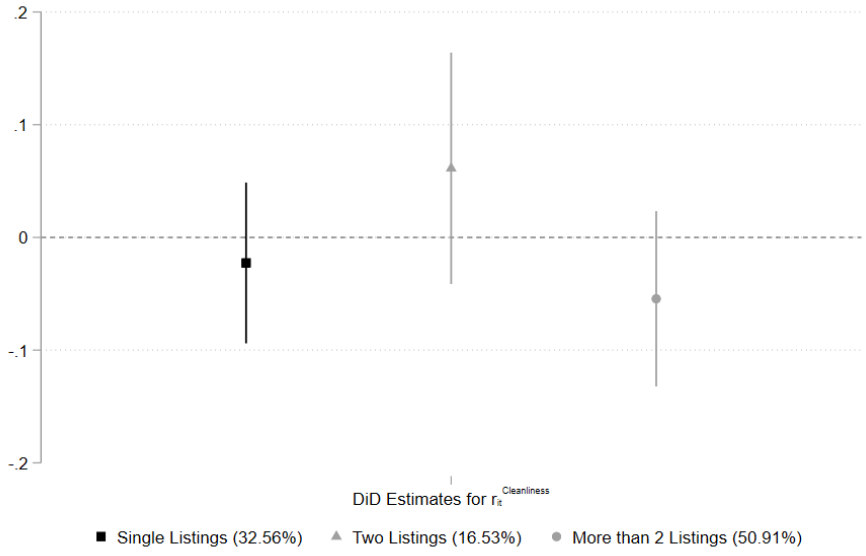


Figure 18: Heterogeneity analysis on ratings on cleanliness by host's number of listings on Airbnb before the HSO approval

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on cleanliness and the estimates' 95% confidence intervals. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the host's number of listings on the platform before the HSO approval.

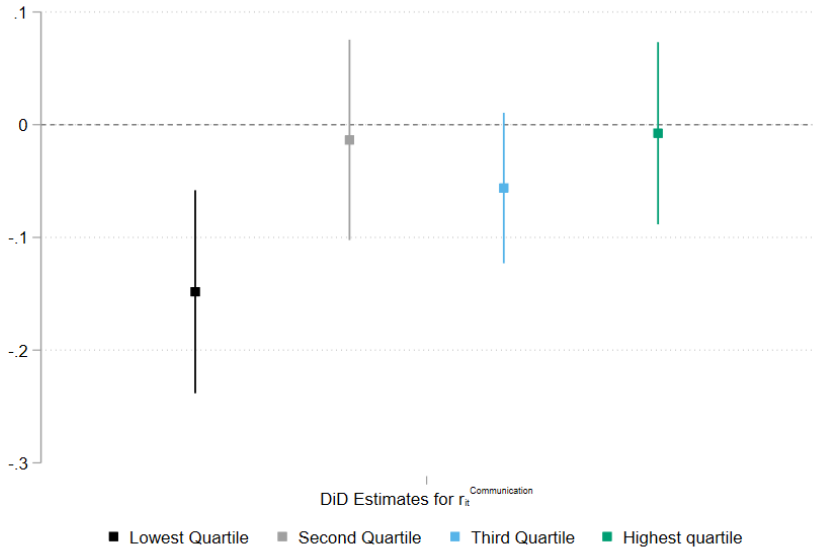


Figure 19: Heterogeneity analysis on ratings on communication by share of owners in the listing's neighbourhood

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on communication and the estimates' 95% confidence intervals. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the share of owners in listing's neighbourhood according to the 2018 census.

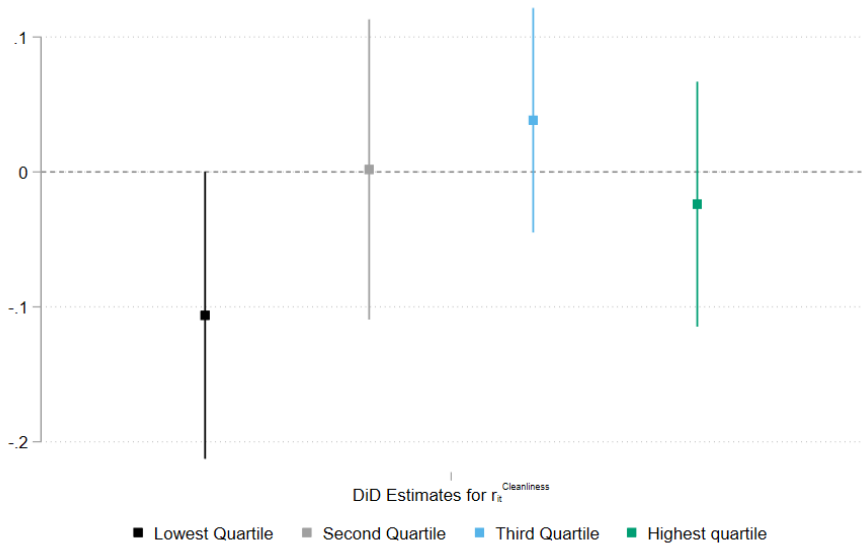


Figure 20: Heterogeneity analysis on ratings on cleanliness by share of owners in the listing's neighbourhood

Notes: The figures plot the estimates of β_2 in Equation 2 for ratings on cleanliness and the estimates' 95% confidence intervals. Number of observations is 57,546. The heterogeneity is performed by splitting the sample given the share of owners in listing's neighbourhood according to the 2018 census.

Additional Results for the Alternative Specification

Table 9: Summary statistics for treated and control listings in alternative specification

	(1) Treated		(2) Control		(3) Difference	
	Mean	S.D.	Mean	S.D.	Δ	p-Value
Listing's tot. reviews ($n_{i,t}$)	28.65	47.07	21.29	37.56	7.36	(0.00)
Listing's n. reviews per month ($n_{i,t} - n_{i,t-1}$)	2.83	2.80	2.51	3.08	0.32	(0.00)
Listing's price per night ($, USD)	143.63	141.25	133.32	177.22	10.31	(0.06)
Overall Rating (#stars)	4.69	0.52	4.68	0.59	0.01	(0.59)
Accuracy Rating (#stars)	4.79	0.54	4.76	0.60	0.03	(0.13)
Check-in Rating (#stars)	4.85	0.48	4.83	0.54	0.01	(0.41)
Cleanliness Rating (#stars)	4.68	0.63	4.68	0.65	0.00	(0.94)
Communication Rating (#stars)	4.86	0.47	4.80	0.61	0.06	(0.00)
Location Rating (#stars)	4.80	0.49	4.75	0.62	0.04	(0.02)
Value-for-money Rating (#stars)	4.69	0.61	4.70	0.63	-0.01	(0.70)
Host's listings (#number)	4.78	9.58	4.52	14.45	0.25	(0.52)
Owners in neighborhood (%)	26.57	22.65	45.90	25.91	-19.33	(0.00)
Number of listings	3,519		931		4,450	

Notes: The table compares the characteristics of treated (column(1)) and control group (column (2)) before the HSO implementation. Column (3) report the difference between the two groups and test its statistical significance with t-test (of which we report the p-Value).

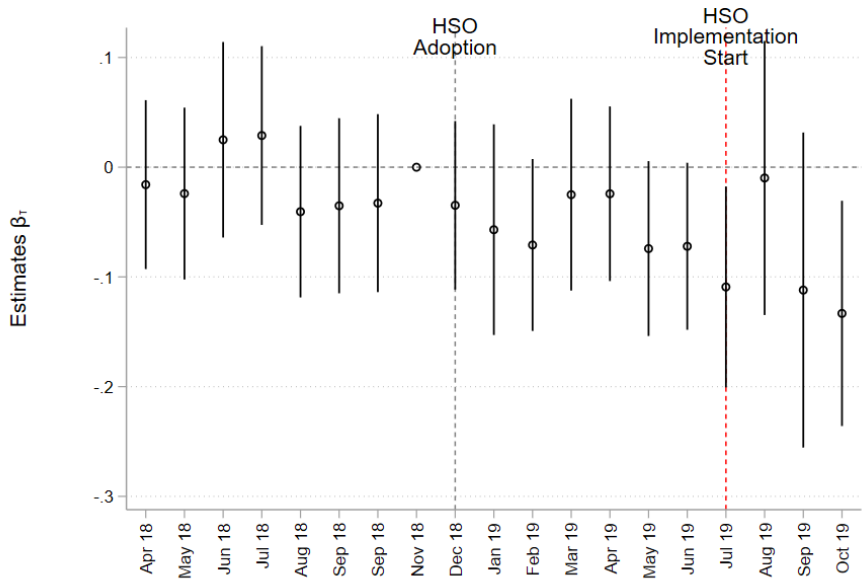


Figure 21: Event study estimates comparing ratings on check-in in the City of Los Angeles with other cities in the county.

Notes: The plot reports the estimates of β_τ from $r_{it}^{check-in} = \sum_{\tau=Apr18}^{Oct19} \beta_\tau LAcity_i \times \mathbb{1}(t=\tau) + \phi X_{it} + \gamma LinearTrend_t + \mu_i + \epsilon_{it}$ and the confidence interval at 95%. Coefficients after the HSO approval become significantly different from the coefficient at the reference period - November 2018 (normalized at zero). We do not find significant pre-trend and the end-game effect is particularly strong after the policy announcement.

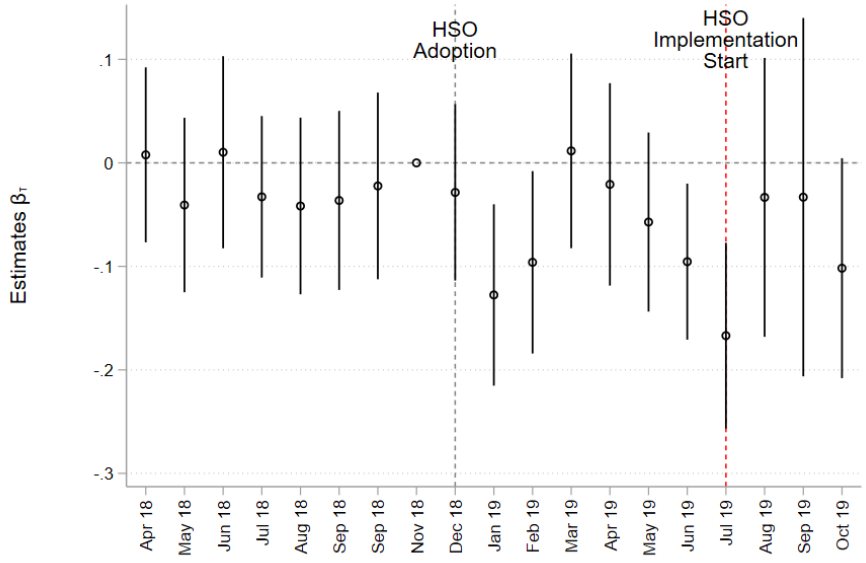


Figure 22: Event study estimates comparing ratings on communication in the City of Los Angeles with other cities in the county.

Notes: The plot reports the estimates of β_τ from $r_{it}^{communication} = \sum_{\tau=Apr18}^{Oct19} \beta_\tau LAcity_i \times \mathbb{1}(t=\tau) + \phi X_{it} + \mu_i + \gamma LinearTrend_t + \epsilon_{it}$ and the confidence interval at 95%. Coefficients after the HSO approval become significantly different from the coefficient at the reference period - November 2018 (normalized at zero). We do not find significant pre-trend and the end-game effect is particularly strong after the policy announcement.

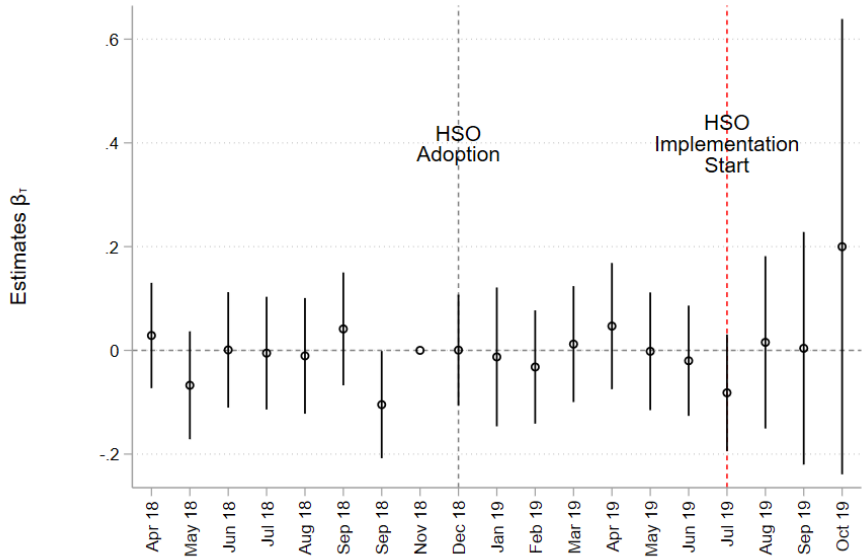


Figure 23: Event study estimates comparing ratings on cleanliness in the City of Los Angeles with other cities in the county.

Notes: The plot reports the estimates of β_τ from $r_{it}^{cleanliness} = \sum_{\tau=Apr18}^{Oct19} \beta_\tau LAcity_i \times \mathbb{1}(t=\tau) + \phi X_{it} + \mu_i\gamma + LinearTrend_t + \epsilon_{it}$ and the confidence interval at 95%. Coefficients after the HSO approval are not significantly different from the coefficient at the reference period - November 2018 (normalized at zero). We find no end-game effect on the listings' ratings on cleanliness.

2 Crowd-sourcing AI Related Tasks: Insights from an Online Labor Platform

This chapter is join work with Ulrich Laitenberger (Tilburg University and Télécom Paris, Institut Polytechnique de Paris) and Paola Tubaro (CREST, Institut Polytechnique de Paris).[23]

2.1 Introduction

The ongoing debate about the impact of Artificial Intelligence (AI) and new technologies on employment primarily focuses on the skilled side of labor demanded by automation (Acemoglu and Autor 2011; Brynjolfsson and Mitchell 2017; Agrawal, Gans, and Goldfarb 2019; Frank et al. 2019; Fossen and Sorgner 2019; Lee and Clarke 2019 and Alekseeva et al. 2021). However, human involvement in AI extends beyond high-skilled labor (e.g., programming), to include the execution of simpler data training, requiring substantial human participation in both generating and annotating data. This "hidden" data work behind AI encompasses tasks like capturing pictures and videos, annotating text and images, transcribing audios, and may even involve testing technologies and impersonating the AI (Tubaro and Casilli 2019; Tubaro, Casilli, and Coville 2020).

Training prediction models accurately requires precision and consistency from annotators, specialized knowledge in semantics and an understanding of cultural and linguistic nuances. Additionally, diversity in the pool of respondents is crucial to ensure a representative outcome in data generation. Obtaining all these elements may not always be possible in-house. Various approaches exist for gathering human inputs for AI, including traditional outsourcing to companies specializing in Business Process Outsourcing (Miceli and Posada 2022; 2021; Le Ludec, Cornet, and Casilli 2023). Alternatively, firms could use crowd-sourcing platforms, such as Amazon Mechanical Turk (AMT) or microWork-

23. This work was supported by the French Research Agency (ANR) under grant ANR-19-CE10-0012 ("HUSH"). We would like to thank microWorkers and especially Nhatvi Nguyen for providing us with the data for this article and their support. We extend our gratitude also to Antonio Casilli and Matthias Hirth. Helpful feedback was received at the INDL-6 conference in Berlin.

ers. These platforms offer a scalable and cost-effective solution to outsource small tasks. Firms can connect to a globally scattered pool of contributors to whom they outsource, or "crowd-source" small or "micro" tasks under a piece-rate compensation (Horton 2010; Hornuf, Mangold, and Yang 2023).[24] Micro-tasks often do not require particular specialized skills and only demand access to an internet connection and an electronic device as equipment. Crowd-sourcing platforms have grown in usage in the past decade (Kässi and Lehdonvirta 2018), with clients ranging from small firms to Fortune 500 companies (Corporaal and Lehdonvirta 2017). Inputs collected on such platforms find application across various domains, from the annotation of medical images and symptoms (Rogstadius et al. 2011; Figueroa et al. 2012; Chandler and Kapelner 2013 and Richter and Khoshgoftaar 2020) to moderating and recognizing hateful text (Arhin et al. 2021; Excell and Moubayed 2021; Larimore et al. 2021; Beck et al. 2022; Davani et al. 2023 and Huang, Kwak, and An 2023).

However, certain characteristics of platform labor may raise concerns regarding its suitability in terms of ethical and privacy-related concerns and ensuring the collection of high-quality work (Thuan, Antunes, and Johnstone 2013; Belletti et al. 2021). First, the data used for annotation may contain sensitive information, and firms may lack confidence in having an anonymous crowd of contributors accessing them. This could discourage the use of crowd-sourced annotation for sensitive data (e.g., medical records). Moreover, ethical concerns may arise from data generation requests, for instance in soliciting the recording of videos of the worker of its closed ones in private spaces. These concerns lie at the heart of the ongoing debate on AI regulation, where notable emphasis is put on requirements on data provenance. Second, on online labor platforms, it may be complex to collect quality contributions given the anonymity of the market. Additionally, the presence of numerous workers contributing to the same project further complicates the selection and monitoring process. This can result in a risk of collecting poorly executed work and being unable to prevent misannotated data from being used in training AI models, leading to biases.

24. Small refers to the limited time it takes to complete the task.

Indeed, while crowd-sourcing data annotation facilitates access a more diverse, and possibly representative pool of annotators, quality of the final outcome is still under debate (Mason and Watts 2010; Shaw, Horton, and Chen 2011). However, the level of risk associated with poor data quality inputs in AI systems depends on the final application of the technology. While errors in training might be less relevant in some contexts, they can prove extremely critical in applications demanding high precision, such as self-driving vehicles (Kretschmer et al. 2023). Beck (2023) extensively summarizes previous interdisciplinary research on the quality aspects of annotated data. Potential biases in data training can, in fact, arise from linguistic perspectives (Beck et al. 2022) or cognitive biases (Eickhoff 2018). Annotators' demographic characteristics can also impact data collection and annotation Al Kuwatly, Wich, and Groh (2020). A large bulk of the literature has focused on and proposed different mechanisms and tools for bias mitigation and better monitoring (Ipeirotis, Provost, and Wang 2010; Hirth, Hoßfeld, and Tran-Gia 2013; Agley et al. 2022 and Rivera et al. 2022). However, evidence about the actual strategies used by crowd-sourcing firms are rarely discussed.

This chapter exploit access to a proprietary dataset of a leading commercial "crowd-sourcing" platform to study the behavior of the demand side of data annotation and generation tasks, emphasizing the significance and attention AI developers allocate to privacy and ethical concerns and, especially, to data training quality. To identify AI related tasks in our data, we employ a text analysis approach, detecting keywords primarily associated with data annotation and generation in the titles and descriptions of jobs. This approach is complemented by a more straightforward identification based on the labels chosen by requesters to categorize launched tasks' campaigns. We explore then the types of tasks that are crowd-sourced in data annotation and generation to understand if ethical or privacy concerns are somehow at play. In a regression setting, we then examine how the behavior of requesters crowd-sourcing AI tasks differs compared to others in terms of strategies to elicit quality of collected tasks. While employers can mitigate such risks by offering higher monetary incentives and investigating quality, platforms also often provide some tools such as the option to target pre-defined groups of workers based

on skills or geographical area, or to delegate the quality investigation of executed labor to the platform itself. In this paper, we explore how these tools are utilized by platform's clients.

Our analysis shows a significant increase, starting in 2019, of demand for data work on the platform. We find that data collection tasks primarily serve market research purposes (e.g., collection of products' prices). In more than half of tasks about data collection and generation, workers are required to generate first-hand information, either by recording themselves, taking pictures or answering to questionnaires. Finally, data annotation tasks often require the identification of emotions and spatial objects for training AI models. Industries dealing with sensitive data are still relatively marginal on the platform, likely indicating privacy concerns of requesters in sharing sensitive data with the "crowd".

Our regression analysis' findings indicate that requesters in the AI domain select more specific worker groups at the campaign design stage by restricting the pool of contributors who can access and execute their tasks. This restriction is primarily achieved through predefined groups based on geographic locations and experience. Our results further reveals that AI related tasks do not offer higher compensations, suggesting that requesters tend to prioritize worker selection over monetary incentives to enhance effort. Finally, we show that tasks associated with data annotation exhibit higher probability of being rejected, indicating a more thorough post-task quality investigation conducted by the requesters. All these results emphasize the importance of the quality of AI execution compared to other tasks. These findings provide valuable insights for newcomers in the market, showcasing the tools they can utilize to ensure quality. Additionally, they inform platforms about the most used tools by requesters that can be strengthened to become more and more attractive, considering the growing relevance of the market of crowd-sourced data work.

The remainder of this paper is organized as follows: in Section 2.2, we describe the microWorkers platform and establish a conceptual framework to formulate hypotheses on how firms crowd-source AI related tasks on the platform. In Section 2.3, we provide an overview of the data used in this study, offering a detailed description of the dataset

and our methodology to detect AI related tasks. In Section 2.4 we present the results, describing first which type of AI related campaigns are launched on the platform. Then, with a regression analysis, we study how AI jobs differ in campaign design and quality investigation. Finally, in Section 2.5 we discuss our findings and conclude.

2.2 Background

2.2.1 microWorkers Platform

microWorkers is a platform connecting employers, hereafter referred as "requesters", with workers who complete small online tasks, known as "micro-jobs" or "micro-tasks".[25] The platform is operated by a US-based company and was launched in 2009. According to the platform website, on the platform it is possible to reach more than 3,750,302 workers worldwide who have been completing 133,094,371 tasks.[26] The relevance of this online labor platform is underscored by the inclusion of microWorkers in various international reports on online labor (Stuart et al. 2017; Berg et al. 2018 and Datta et al. 2023).

On microWorkers, users can register from any location. Their user's accounts allow them to request and conduct tasks. Requesters initiate a campaign consisting of a batch of small tasks, typically compensated between 0.10 and a few dollars (Hirth, Hoßfeld, and Tran-Gia 2011). To design a campaign, requesters can decide to use one of the templates provided by the platform or design tasks from scratch. According to the content of the tasks to outsource, requesters select a category to label their campaign, add a campaign's title and a description of the task's content. The length of task's instruction can be modulated by the employer to increase the clarity of requirements. Table 17 in Appendix 2.6 shows the large scope of the platforms, with tasks ranging from marketing and online promotion (e.g., Search Engine Optimization - SEO, offers, and sign-ups), interaction with social media content, but also data annotation, research surveys (e.g., answering surveys), Q&A, etc.

While payment per task is set by the outsourcing firm at campaign level (i.e. each

25. We use the term' 'requesters" to identify the demand side of the platform for consistency with the extensive literature on Amazon Mechanical Turk, where employers are indeed referred with this word.
26. Last visualization of https://www.microworkers.com/ on the 30/01/2024.

task in the same campaign is paid the same), each job category is associated with different minimum wage depending on the job type and the required worker's location, in cases of targeted demand. Indeed when launching a campaign, requesters can choose either to leave it open to all the contributors active on the platform or can decide to restrict the pool of contributors who can access it by targeting specific groups of workers based on country of origin, macro geographic area, or through a so called "hiring group" mechanism. The latter option allow to hire via predefined workers' groups created either by the platform (e.g., selecting the best annotator or workers from a specific location) or by the requester itself (e.g., choosing workers from previous collaborations or cherry-picked workers by features displayed on the platform). Requesters also indicate the expected time it should take a worker to finish a task and decide the size of the campaign by setting the number of tasks in a campaign (available positions), and the maximum number of tasks that can be done by the same worker within a campaign. At the design stage, requesters can also decide whether to enable the "auto-rating" option to delegate the platform's operator to assess the submitted work's quality, under extra charge. To require this service, the outsourcing firm has to provide the platform which guidance on accuracy, adherence to instructions, or other task-specific requirements to guide the task evaluation.

Campaigns must be approved by the platform before being launched. Once approved, open campaigns are made available on the platform. All tasks available to a worker are visible on the platform's interface (Figure 24) along with payment details and other task dimensions, such as the expected time to complete the job and the time for the requester to rate the task and pay the worker. Workers self-select tasks that are available to them and execute them, with no ex ante approval required from the employer and no bargaining involved. Tasks payment is indeed a "take it or leave it" offer.

Once completed, tasks are submitted for validation and are either validated or rejected. Workers are paid only when a task is successfully rated, either because the requester approves and validates the output or automatically after 7 days if the task remains unreviewed. Requesters can reject a task but must provide the platform with a written explanation. The platform implements a reputation system in which workers are eval-

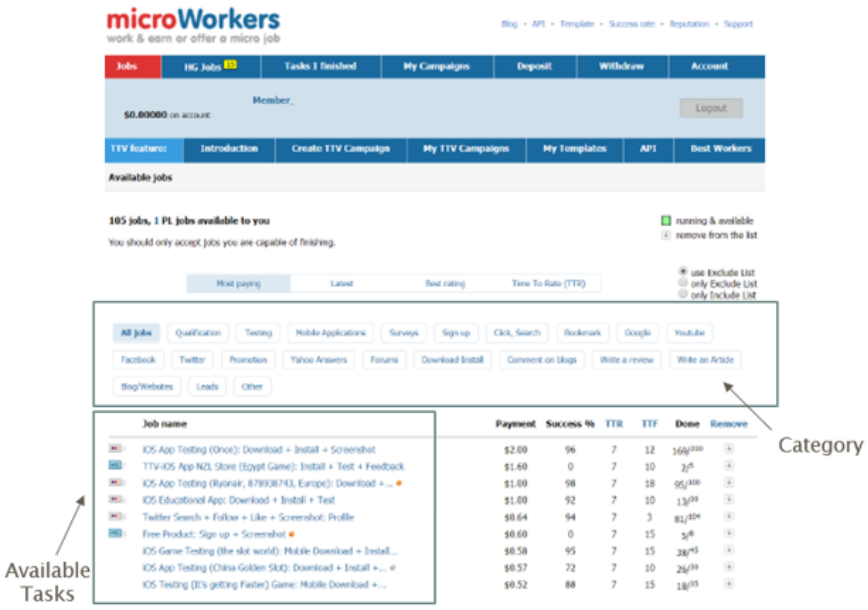


Figure 24: microWorkers interface

Notes: Screenshot of the microWorkers interface. The image was taken in July 2022.

uated on the basis of their success rate, namely the share of tasks that they complete successfully. To continue executing tasks on the platform, workers must maintain a success rate of at least 75%. However, some tasks may be complex to investigate and a lack of assessments may lead then to inflated reputation scores that become not particularly informative.

While the platform has been existing since 2009 already, it has been further growing during the Covid-19 Pandemic. We provide descriptive figures about recent trends on the platform in the Appendix 2.6.

2.2.2 Crowd-sourcing AI related tasks: Conceptual Framework

In this sub-section, we design a conceptual framework to clearly define hypotheses on the specificity of AI related demand that will guide our platform analysis.

Firms willing to outsource AI related tasks to the crowd of workers active on online platforms face different challenges. First, data annotation tasks often involve sensitive data, raising concerns about privacy, and ethical considerations. Crowd-sourcing may not be applicable in all domains due to the nature of the data involved. Therefore, sharing

sensitive information with a crowd of anonymous individuals may not be ideal for some privacy-concerned firms. In this case, using this channel to outsource data work may be chosen only for tasks with less sensitive content, avoiding, for instance, crowd-sourcing tasks that involve handling medical information (**Hypothesis 1**).

Once decided to crowd-source a task, requesters may be concerned by how to ensure quality of the executed tasks. The quality of task execution is of crucial importance in AI related micro-tasks. Inadequate execution during annotation and data generation poses a risk of introducing bias when training models. Therefore, a thoughtful and effective strategy from outsourcing firms is needed to i) ensure the quality of the execution and ii) be able to detect and reject mistakes. To achieve the first goal, requesters could exogenously select the worker that better fit their need by targeting specific contributors. If this is the case, we would expect AI to record differences in hiring patterns compared to other sectors, with higher targeting based on geographical considerations (specific geographic areas if specific language and cultural skills are required for the job) or the use of "hiring groups" (**Hypothesis 2a**). Requesters of AI related tasks can also (or alternatively) set incentives to attract better workers and incentivize their effort, with the most straightforward being monetary incentives. We may therefore expect higher AI task to have higher pay rates (**Hypothesis 2b**).

At the campaign design stage, firms can also opt for delegation to the platform of the quality investigation via the "auto-rated" mode. This serves to achieve the second goal, the investigation of task outcome to filter poor quality. Delegation to the platform for monitoring quality via the "auto-rating" mode incurs an extra cost, and firms are required to provide microWorkers with specific guidelines to guide the investigation of the quality of executed work. We test if AI related campaigns more often utilize this option, in line with the larger relevance of the quality of execution in the domain **(Hypothesis 3)**.

After a task is completed and submitted by the worker, requesters take another important strategic decision. They can choose whether to investigate or not the quality to filter and reject poorly executed tasks. Yet, implementing a robust monitoring system incurs costs and may counterbalance the advantage of accessing cheap labor. It may be

worthwhile only if poor quality has very adverse consequences, as may be the case with misannotated data. We expect requesters of AI campaigns to investigate quality more in order to avoid introducing compromised data in AI training (**Hypothesis 4**).

2.3 Data

We access proprietary data from the crowd-sourcing platform microWorkers, spanning from 2016 to the first quarter of 2021. This dataset includes task details such as the task's category, payment of the campaign, job description, and the validation/rejection outcome for each task. We also have information about users' experiences on the platform, including entry dates, completed tasks (for the worker), launched campaigns (for the requester), and geographical location. We merge the different databases at the campaign, task, and user levels to obtain a comprehensive dataset. It is important to note that the value of our unique dataset comes from the fact that it is rare to have access to such comprehensive data, offering an opportunity to observe aspects that are usually obscured from view or scraped data, such as the validation/rejection outcome of a task.

2.3.1 Identification of AI Related Tasks

To identify AI related tasks, we start by looking at the campaign identification system implemented by the platform that includes a set of labeled categories the requester could choose for characterizing the job offer. We classify as an AI related jobs, those tasks belonging to a campaign assigned to one of the following categories: "Data Collection/ Mining/ Extraction/ AI Training", "Data Annotation" and "Data Transcription". Yet, by solely examining labeled categories, we risk to underestimate the overall size of AI related campaigns on the platform. This is because new categories emerge over time to address evolving AI frontiers. Additionally, tasks requiring human input for AI training or verification may be hidden behind other labels or wording. For instance, the requester may choose not to use AI or data related labels to attract a wider pool of contributors or to avoid "bias" in certain processes (e.g., for a task that requires workers to take a picture of themselves, the requester may choose not to specify the final application of

their job).

To better identify all the tasks that could enter the AI value chain, we follow the approach of Duch-Brown, Gomez-Herrera, et al. (2022) and use a set of keywords related to AI and data work to identify AI related campaigns from the job title and description, in addition to the official label. In order to compile the list of words that are closely associated with AI and data training, we rely on descriptions found in the literature and use our first-hand qualitative research experience. Unlike Duch-Brown, Gomez-Herrera, et al. (2022), who study the freelancing market for skilled workers and focus on technical vocabulary related to automation, our keywords refer to simple tasks, mostly in data annotation (e.g., "label", "transcribe") and data generation (e.g., "video", "image"). In Appendix 2.6, we report the full list of keywords we used.

Figure 25 shows that the demand for AI related campaigns has greatly increased starting from 2019 to the point that, in the first quarter of 2021, the number of launched campaigns whose category label was somehow related to AI was almost 140 times larger than in the first quarter of 2016. On the other hand, the graph illustrates a flat evolution for the demand of non-AI related campaigns. At the same time, there is no growth for non-labeled AI tasks (detected with our text analysis algorithm), but even a small relative decline is observed, likely due to the availability of new AI related labels on the platform. Indeed, the first campaign labeled as "Data Transcription" only dates back to January 2020, and the category "Data Annotation" makes its first appearance in July 2020.

Figure 30 in Appendix 2.6 shows that among the most frequent keywords on our list detected in"hidden" AI campaigns, there is the word "data", words related to data generation such as "record"," camera", "resolution" and to data annotation, such as for example: "label", "categories", "row", "cell", "column", etc. Table 18 in Appendix 2.6 presents the frequency of this unlabeled AI work in the different campaign categories. We see a similar distribution of the overall distribution of campaigns shown in the table, with the majority of hidden AI being labeled in the popular categories "SEO & web traffic", "video music sharing platform" and "offer and sign up".

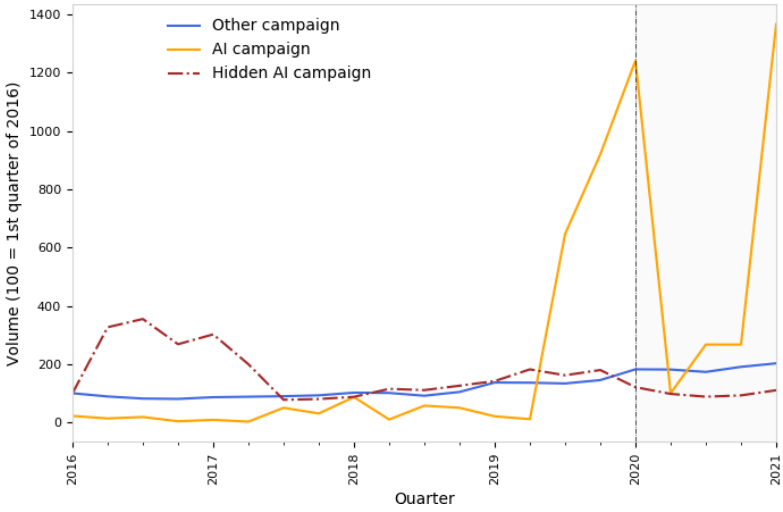


Figure 25: Relative growth of demand by campaign type (AI and non-AI)

Notes: The figure illustrates the quarterly evolution of the number of campaigns launched on the platform for various categories from January 2016. Values are normalized with the first 2016 as the reference period set at 100. The shadowed area represent the period after January 2020, and points to the beginning of the Covid-19 pandemic. The yellow line track demand for campaigns labeled as "Data Annotation", "Data Transcription" and "Data Collection/ Mining/ Extraction/ AI training". It shows that the demand for AI related campaigns has greatly increased starting from 2019 to the point that, in the first quarter of 2021, the number of launched campaigns whose category label was somehow related to AI was almost 140 times larger than in the first quarter of 2016. The dashed red line tracks the execution of campaigns not initially labeled as AI but later assigned to AI after passing through our text analysis algorithm. There is no growth for non-labeled AI tasks (detected with our text analysis algorithm), but even a small relative decline is observed. Finally, the blue line represents the flatter evolution remaining non-AI related campaigns (even if a growing trend is detected from 2019).

2.3.2 Analytical Sample

For our regression analysis, we narrow down the sample by focusing on tasks executed between July 2020 and the end of April 2021. We select July 2020 as it marks the moment when all category labels identified as AI related exist on the platform. Our analytical sample comprises 6,422,752 tasks executed by 10,992 workers, collected in 62,944 campaigns initiated by 2,367 requesters.

In the data we observe several characteristics of the campaign (and the task) that we use to build our main variables at campaign (and task) level:

- AI related labelled campaign: Binary variable (dummy) which identifies if a campaign is labeled with one of the categories we identified as AI-related. It takes a value of 1 if the campaign belongs to the categories: "Data Collection/ Mining,

etc.", and "Data Transcription", 0 otherwise.[27]

- AI related labelled campaign + "hidden" AI: Binary variable that takes a value of 1 if the campaign is labeled in one AI-related category. Additionally, it also takes a value of 1 if the campaign belongs to a different category but was detected containing some AI keywords using our text analysis approach (i.e."hidden" AI campaign).

- Data Annotation, Data Collection/ Mining/ etc and Data Transcription: Set of mutually exclusive binary variables, taking value 1 if a campaign belongs to that category, 0 otherwise.

- "Hiring-Group" campaign: Binary variable that takes value 1 if the campaign is directed to a specific pre-selected group of contributors, called "hiring group", 0 otherwise.

- Non-Targeted: Binary variable that takes value 1 if a campaign is open to all workers available on the platform, regardless skills, experience and location. It takes value zero otherwise. Alternatively a campaign can be directed to geographical areas. We aggregate two of the most used geo-target in two clusters: Target West (for campaigns targeting workers in North America, Australia, New Zealand and Western Europe) and Target Asia.[28]

- "Auto-Rated": Binary variable that indicate if campaign uses the "auto-rated" option (i.e. pay an extra to outsource the verification of tasks execution quality to the platform). In this case the variable takes value 1, otherwise it takes value 0.

- Payment per task: Continuous variable measuring the monetary compensation assigned to the worker if the task is validated. The variation of this variable occurs at the campaign level, meaning that within a specific campaign, all tasks receive

27. We re-scaled the dummy variables, traditionally 0/1, as 0 or 100 binary variable so that the mean value represents directly the share of campaigns that are AI-related.

28. Note that "hiring group", non-targeted, target West and target Asia are mutually exclusive possibilities. There are additionally some campaigns that do not belong to any of these categories since they target workers in other geographical areas (e.g., Africa, Eastern Europe)

the same amount of compensation. Firms wage is established at the launch of the campaign and is visible to workers on the platform interface. We drop from the sample tasks with zero wage, such as qualification tests.

- Expected Execution Time: This continuous variable provides insight into the employer's initial assessment of the minutes required for task completion. It is important to note that this variable does not measure the realized execution time, which is not observed. The expected execution time vary across, but not within, campaign.

- Length instructions: Continuous variable built by counting the number of words used to describe the task. It serves as a proxy for the level of detail in the description, but also indicates the time workers need to invest in understanding what is required of them.

- Task is rejected: Binary variable at task level. It takes value of 1 if a task faces rejection and 0 if it does not. A rejected task reflects the employer's dissatisfaction with the quality of executed work. The variable mean represents the share of tasks that were rejected. We remove from the sample tasks with pending rating status (i.e. the task is not yet validated or rejected).

We then utilize the entire history of platform's data to construct time-dynamic variables at user level. Time-varying firm's characteristics were collected at the moment of the first tasks executed in a new campaign. Conversely, data at the worker level are updated task after task.

- Days since sign-up: Continuous variable which counts the days since the user's registration on the platform.

- Tot. campaigns launched: Continuous variable which counts the campaigns launched by a requester since the firm's registration on the platform.

- Tot. task executed: Continuous variable which counts the tasks executed (and rated) by a worker since registration on the platform.

- Success rate: Variable measuring the share of tasks executed by a worker that were validated and, thus, paid.

- Finally, we integrate platform data with administrative data on the GDP per capita of 2020 (in USD) in the user's country from the International Monetary Fund (IMF). Datasets are merged based on the information on the location as self-declared by the user. We drop in the country of location of the user was not available.

Table 10: Distribution of the main variables in the analytical sample

	N.	Mean	S.D.	Min	Max
Panel A1: Campaign Level					
AI related labelled campaign (binary)	62,944	2.512	15.648	0	100
AI related labelled campaign + "hidden" AI (binary)	62,944	6.067	23.873	0	100
Data annotation (binary)	62,944	1.932	13.764	0	100
Data collection/mining/extraction/AI training (binary)	62,944	0.574	7.551	0	100
Data transcription (binary)	62,944	0.006	0.797	0	100
Hiring Group campaign (binary)	62,944	29.518	45.613	0	100
Non-targeted (binary)	62,944	43.073	49.518	0	100
Target West (binary)	62,944	19.578	39.680	0	100
Target Asia (binary)	62,944	5.065	21.928	0	100
Auto-Rated campaign (binary)	62,944	0.229	4.778	0	100
Payment per task ($, USD)	62,944	0.187	0.280	0.020	20
Expected task duration (#minutes to finish)	62,944	7.597	19.980	1	3,600
Campaign size (#available positions)	62,944	125.720	359.671	1	55,123
Length instructions (#words)	62,944	43.984	60.944	2	2,424
Panel A2: Requester-Campaign Level					
Days since sign-up	62,944	1,242.354	1,094.350	0	4,330
Tot. campaigns launched	62,944	1,820.596	3,625.644	1	31,514
GDP per capita country (2020)	62,944	483.905	258.355	7.099	1000
Panel B1: Task Level					
Task is rejected (dummy)	6,422,752	2.766	16.399	0	100
Panel B2: Worker-Task Level					
Tot. tasks executed	6,422,752	3782.453	7699.089	1	8868
Days since sign-up	6,422,752	637.743	871.935	0	4.317
Success rate	6,422,752	96.952	8.303	0	100
GDP per capita country (2020)	6,422,752	76.269	143.435	2.210	1000

Note: The table summarizes the main moments of the distribution (number of observations, mean, standard deviation, minimum and maximum value) of the variables used in the analysis in the analytical sample, including tasks in campaign executed between July 2020 and April 2021. Panel A shows the distribution of variables at the campaign level, Panel A2 at the campaign-requester level. Panel B presents variables at the task level, and B2 at the worker-task level. All data come from the platform, with the exception of GDP per capita, collected from IMF data for 2020 and measured in USD. The GDP value has been rescaled to a range with a maximum of 1000. Binary variables are rescaled to vary between 0 and 100, to be able to read the average as the share of observations where that variable has realization 1 (or, equivalently 100). For example, the first row of Panel A1 shows that around 2.5% of camapigns launched on the platform between July 2020 and April 2021 were labelled in AI related category.

Figure 10 provides descriptive statistics for the main variables in the analysis. Panel

A describes the distribution of variables at campaign level. Campaigns labeled with in AI related category represent almost 3% of all campaigns whose tasks were executed between July 2020 and April 2021, with tasks labelled as "Data Annotation" accounting for almost 2% of all campaigns and those labelled in the category "Data Collection/ Mining/ Extraction/ AI Training" for less than 1%. There is also a negligible number of tasks under the category "Data Transcription". The share of AI campaigns increases to almost 6% if we take into account the "hidden" AI related campaigns. Almost 30% of tasks are assigned through a "hiring group", yet, more than 43% of campaign did not filter the contributors that could perform their tasks. Yet, when present, the major geographical target are workers located in the "West" macro-area. This demand accounts for almost 20% of all campaigns and it is followed by demand targeted to workers located in Asia (5%) and a negligible shares of workers from other geographical areas. This results resonates with the fact that the largest bulk of the demand on microWorkers comes from North America and it is possible that these requester aim at targeting English-speaking countries or workers with similar cultural backgrounds.[29] The "auto-rated" option is used only in a negligible share of campaigns. On average, tasks on the platform are compensated at approximately $0.19 USD. Requesters estimate that it takes slightly less than 8 minutes for a task to be completed. Campaigns vary in size from 1 to 55,123 tasks, and contain on average 126 tasks. A mean of around 44 words are used to describe the task and explain the job.

Panel A2 summarize the distribution of variables at requester-campaign level. Requesters time on the platform at the campaign launch spans from 0 to 12 years. Yet the majority of campaigns are launched in the first 3.4 years on the platform of the requester. On average, requesters launch 1,820 campaigns. Panel B illustrates the distribution of variables at task level. It shows a limited rejection rate on platform. Out of 6,422,752 tasks, less than 3% are rejected and therefore not paid by the firm or individual requester, meaning that the rest are either organically approved or automatically validated after the 7-days deadline with no investigation by the requester.

29. Table 29 in Appendix 2.6 illustrates the evolution of the share of requesters on the platform by location.

Panels B2 describe worker-task level variables. The time-variant data about users are updated task after task. Workers stay less on the platform compared to requesters, as indicated by the fact that average task is executed by workers who entered the platform since slightly more than 1 year. Workers perform on average almost 3,800 tasks, but with large variation across users. Yet the very high average success rate (97%) is pretty homogeneous across worker-tasks observation and reflects the low rejection rate in Panel B. Finally, by looking at the distribution of GDP per capita, we show that outsourcing firms or individuals are located in richer countries than platform workers.

2.4 Results

2.4.1 Analysis and Results and the Type of AI Campaigns

We begin our analysis by examining tasks categorized as "Data Collection/ Mining/ Extraction/ AI Training". We identify the most frequently occurring words in the campaign titles and descriptions. We distinguish vocabulary related to the required actions from words indicating industries or application and scopes of the final output. We manually cluster words into different groups and report the frequency of each group in Table 11.[30]

Panel A show the occurrence of keywords clusters related to required actions. We are able to identify almost 61% of campaigns using the words we clustered. About 55% of campaign in category "Data Collection/ Mining/ Extraction/ AI Training" involve the generation and collection of organic data, meaning generation of firsthand data by recording videos, audios, taking pictures, and providing information through questionnaires. 35% of these campaigns require workers to collect and report data either online and, less often, offline. Other tasks in this category include solving captchas, and negligible share of campaigns involving activities such as testing, training and moderation activities.

Panel B reports the occurrence of keyword clusters concerning the industry application for which the data are collected/generated. We are able to detect almost 62% of

30. Tables 19 and 20 in Appendix 2.6 list the keywords included in each cluster of words used in Tables 11 and 12.

Table 11: Occurrence of words from thematic clusters in title and description of campaigns in the "Data Collection/ Mining/ Extraction/ AI Training" category

Cluster of words	Occurrence in campaigns
Panel A: Required Actions	
Generating Data & Providing Data	54.56%
Collecting Data & Detecting/Annotating	35.40%
Solving	1.46%
Testing	1.16%
Training	0.55%
Moderating	0.06%
Total	60.77%
Panel B: Industries/Applications	
Real estate, finance	38.38%
Vehicles	33.64%
Maps/Spatial Data	12.56%
Sales, promotions, marketing	9.98%
AI, Captchas, Sound and Image Recognition	6.63%
Fitness	4.93%
Search and Engine Optimization	3.89%
Video Games	1.82%
Grocery, Supermarkets	1.34%
Social Media & Blogs	1.22%
Contacts collections	1.09%
Apps	1.03%
Restaurants	0.49%
Videos, Movies, Books	0.48%
Health & Covid19	0.42%
Human Resources	0.24%
Riddles	0.24%
Research, Studies	0.24%
Weather, Nature	0.18%
Total	61.74%

Notes: The table summarizes the occurrence of words clustered into different actions (Panel A) and industry or final work application (Panel B) in "Data Collection/ Mining/ Extraction/ AI Training" campaigns from January 2016 to April 2021. Keywords within each cluster are available in Table 19 in Appendix 2.6. The count is based on the presence of at least one word in a cluster in the title and description. We only focus on English vocabulary and excluded from the list words with occurring only once, auxiliary and modal verbs, adjectives, adverbs, number, blurry words and connectors. Clustering was performed manually, with consistency verified for ambiguous cases. While there are no overlapping words across clusters, a campaign may be counted in more than one category. Total accounts for the share of campaigns who have at least one of all the words we identify in title or description.

campaigns using our list of words. The primary final application is primarily related to market research in the domains of financial and real estate information (around 38%), followed by vehicles (33%). Approximately 12% pertains to maps and spatial information. Data generation may require video and can involve some personal information, while data collection sometimes concerns sectors that may appear more sensitive, such as the medical healthcare sector. However, in this case, it is not about worker information

but rather market research on practitioners.

Table 12 focuses on the "Data Annotation" category and reports the occurrence of words cluster in terms of actions required (Panel A), format of the file to annotate (Panel B) and main objects of the annotation (Panel C) in campaigns' title and instructions.

Table 12: Occurrence of words from thematic clusters in title and description of campaigns in the "Data Annotation" category

Cluster of words	Occurrence in campaigns
Panel A: Required Actions	
Select, Search, Check	1.49%
Annotate	5.51%
Evaluate Classify Order	0.39%
Add, Type, Draw	1.57%
Answer	36.98%
Listen, Watch	4.01%
Suggest and Promote	1.57%
Total	36.98%
Panel B: Input Format	
Image	49.41%
Video	14.32%
Audio	10.78%
Total	49.41%
Panel C: Industries/Applications	
Promotions	49.25%
Facial expressions detection	0.31%
Cartoons	0.39%
Maps and Aerial Pictures	0.16%
Total	49.88%

Notes: The table summarizes the occurrence of words clustered into different actions (Panel A), file of the format to annotate (Panel B) and industry or final work application (Panel C) in "Data Annotation" campaigns from January 2016 to April 2021. Keywords within each cluster are available in Table 20 in Appendix 2.6. The count is based on the presence of at least one word in a cluster in the title and description. We only focus on English vocabulary and excluded from the list words with occurring only once, auxiliary and modal verbs, adjectives, adverbs, number, blurry words and connectors. Clustering was performed manually, with consistency verified for ambiguous cases. while there are no overlapping words across clusters, a campaign may be counted in more than one category. Total accounts for the share of campaigns who have at least one of all the words we identify in title or description.

Panel A can detect around 23.4% of tasks, this partly because many times the description includes links and no textual information that can be used with our techniques. The largest portion is related to answering questions after visualizing a video or some other online content. In Panel B, we show that almost half of the data annotation demand comprises image annotation, with video and audio annotation accounting for 14% and 10%, respectively. Panel C again shows that most tasks are related to promotional

activities. While we cannot detect the source of the image used for annotation (e.g., in facial expression detection), we do not find tasks entering sensitive spheres, such as the medical sector likely due to firms' privacy concerns.

2.4.2 Preliminary Evidence on Requesters' Selection and Moderation Decisions

We begin the analysis on the requesters crowd-sourcing decisions by examining the differences between AI and non-AI tasks from raw data. Table 13 summarizes and compares the means of the main variables of interest between campaigns that we identify as AI related (either by label or with text analysis) and other campaigns.

Table 13: Distribution of main variables according to the AI nature of the campaign/task

	(1) AI related		(2) Non-AI related		(3) Difference	
	Mean	S.D.	Mean	S.D.	Δ	p-Value
Panel A: Campaign Level						
Auto-Rated campaign (binary)	2.99	17.02	0.05	2.25	2.93	(0.00)
Hiring Group campaign (binary)	46.66	49.89	28.41	45.10	18.25	(0.00)
No target (binary)	19.25	39.43	44.61	49.71	-25.37	(0.00)
Target Asia (binary)	5.29	22.39	5.05	21.90	0.24	(0.51)
Target West (binary)	25.66	43.68	19.18	39.38	6.48	(0.00)
Payment per task ($, USD)	0.24	0.34	0.18	0.27	0.06	(0.00)
Expected task duration (#minutes to finish)	15.87	67.42	7.06	11.26	8.80	(0.00)
Campaign size (#available positions)	90.84	960.02	127.97	279.50	-37.13	(0.00)
Length instructions (#words)	76.07	108.90	41.91	55.83	34.16	(0.00)
Number of Observations	3,819		59,125		62,944	
Panel B: Task Level						
Task is rejected (binary)	4.16	19.98	2.70	16.22	1.46	(0.00)
Number of Observations	3,277,916		6,144,836		6,422,752	

Notes: The table presents the mean value and standard deviation for the main variables in the analysis based on the type of campaign launched. As in Table10 the dummy variables are rescaled between 0 and 100 so that the mean reads as the share of observations where the variable takes value 1. Column (1), focuses on AI related jobs, including tasks labeled as "Data Annotation", "Data Transcription" and "Data Collection/ Mining/ Extraction/ AI Training" and those detected with text analysis. Column (2) covers non-AI campaign. Column (3) indicates the difference between the two groups and tests how significant it is from zero. Panel A pertains to campaign-level data, while Panel B concerns rejection at the task level. The table shows statistically significant differences in almost all variables with the exception of the share of the demand targeted to Asian workers.

Panel A focus on the variables set by requesters at the campaign design stage. First, we show that almost all the demand for "auto-rated" campaigns concerns AI tasks. This can be interpreted as preliminary evidence of a larger interest in skimming poor-quality

contributions in AI tasks. Concerning targeting, workers from a specific geographical area or recruited via a "hiring group" are more frequently employed in data training compared to non AI related jobs. As a matter of fact, the demand for contributors located in Western countries is almost 7 percentage point larger when a campaign has to do with AI. On the other hand, there is no significant difference in the share of workers from Asia in AI and non-AI campaigns. Payment per task is on average $0.06 USD higher for AI campaigns and AI tasks are expected to last almost 9 minutes longer, so likely being more complex than other jobs on the platform. Yet, the latter are smaller in size, with an average of around 37 tasks less than the non-AI, despite longer description of the job (detailed with around 34 words more compared to non-AI tasks). Finally, at the task level, the rejection rate for AI tasks appears larger at first glance compared to non-AI tasks.

It is important to notice that various factors might influence the covriates. To better isolate and understand the role of the AI nature of campaigns, we proceed with a regression framework.

2.4.3 Regression Analysis on Campaign Design

We start our regression analysis by exploring the difference between AI and non-AI campaign design in terms of workers selection by the requester and the decision to use the "auto-rated" service. We estimate the following econometric model for each campaign j launched by requester e at time t:

$$Y_{j,t}^{ex-ante} = \alpha + \beta_1 Annotate_j + \beta_2 Generate_j + \beta_3 Hidden_j + \sigma_1' X_j' + \sigma_2' Z_{e,t}' + \phi_t + target_j + \epsilon_{j,e,t} \tag{5}$$

Where the dependent variable $Y_{j,t}^{ex-ante}$ represents if the campaign uses the "Auto-Rated" option or the different targeting options that the requester can use to filter workers (i.e. "hiring group", geographic targeting). We regress $Y_{j,t}^{ex-ante}$ over a set of dummies indicating whether the campaign has some manifest or "hidden" AI content. $Annotate_j$

takes value 1 if the campaign is labelled as "Data annotation and transcription" or "'Data Transcription". $Generate_j$ takes value 1 if the campaign belongs to "Data Collection/ Mining/ Extraction/AI Training" category while $Hidden_j$ indicates the additional AI related tasks detected with text analysis of campaign's title and description. The dummies compare to a situation (dummy=0) where the campaign has no AI related content. We incorporate a range of controls at both the task (X_j') and requester-creation day level ($Z_{e,t}'$). Furthermore we include a fixed effects for the campaign creation day (ϕ_t). We cluster standard errors at the requester level.

Table 14: Coefficient estimates from Equation 5 for "auto-rated" usage and targeting

	(1) "Auto-Rated"	(2) "Hiring-Group"	(3) No Target
Data annotation and transcription ($Annotate_j$=1)	0.091	0.685***	-0.465***
	(0.089)	(0.085)	(0.121)
Data Collection/Mining/Extraction/AI training ($Generate_j$=1)	0.096	0.498***	-0.504***
	(0.066)	(0.133)	(0.076)
"Hidden AI" ($Hidden_j$=1)	-0.002	-0.040	-0.061
	(0.002)	(0.097)	(0.131)
Log Expected task duration (#minutes to finish)	-0.002	-0.189***	0.178***
	(0.001)	(0.069)	(0.062)
Log Campaign size (#available positions)	0.004	-0.100***	0.133***
	(0.002)	(0.017)	(0.021)
Log Length instructions (#words)	0.002*	-0.086**	-0.023
	(0.001)	(0.036)	(0.049)
Log Days since requester's sign-up	0.001*	0.045***	-0.003
	(0.001)	(0.015)	(0.023)
Log Tot. campaigns launched requester	-0.001**	0.018	-0.001
	(0.001)	(0.013)	(0.025)
Log GDP per capita requester's country (2020)	0.000	-0.108***	0.038
	(0.000)	(0.020)	(0.026)
Constant	-0.020*	1.548***	-0.543*
	(0.012)	(0.218)	(0.315)
Launch Day FE	✓	✓	✓
Standard Errors Clustering Level	Requester	Requester	Requester
R^2	0.12	0.39	0.24
Number of observations	62,944	62,944	62,944

* p¡0.1, ** p¡0.05, *** p¡0.01

Notes: The table reports the coefficients estimates from Equation 5. Standards error clustered at employer level are in parentheses. Columns (3) and (4) do not control for target fixed effect since collinear with the main dependent variables, measuring indeed target or non-target of demand.

Estimates of Equation 5 are reported in Table 14. Results in column (1) show no statistically significant difference in the use of the "auto-rated" option once controlled for several confounding factors at the campaign and requester levels. Yet, results of columns (2) and (3) indicate that the employers for AI jobs make significantly more use of the platform's feature to target specific workers. Indeed, AI demand exhibit noticeable lower

share of demand open to all workers with no filters and a larger use of hiring-group tasks in data annotation and generation categories.[31]

2.4.4 Regression Analysis on Wage Setting

We now test whether AI related campaign offer larger monetary incentives than others campaign to attract quality contributions and stimulate workers' effort. We also investigate if campaigns features set at campaign design stage affect the price decision. We are estimating Equation 5, taking as $Y_{j,t}^{ex-ante}$ the logarithm of payment per task.

Table 15: Coefficient estimates from Equation 5 for wage setting

	(1) Log Payment per Task	(2) Log Payment per Task	(3) Log Payment per Task
Data annotation and transcription ($Annotate_j$ =1)	0.078***	-0.029	-0.033
	(0.023)	(0.033)	(0.027)
Data Collection/Mining/Extraction/AI training ($Generate_j$ = 1)	0.104	0.049	0.029
	(0.064)	(0.044)	(0.048)
"Hidden AI" ($Hidden_j$ =1)	0.019	-0.027	-0.037**
	(0.030)	(0.023)	(0.018)
Log Expected task duration (#minutes to finish)		0.050***	0.059***
		(0.014)	(0.012)
Log Campaign size (#available positions)		-0.037***	-0.027***
		(0.006)	(0.005)
Log Length instructions (#words)		0.036***	0.028***
		(0.012)	(0.007)
Log Days since requester's sign-up		-0.003	-0.000
		(0.005)	(0.003)
Log Tot. campaigns launched requester		-0.012	-0.011**
		(0.007)	(0.005)
Log GDP per capita requester's country (2020)		0.020***	0.018***
		(0.006)	(0.004)
"Auto-Rated" campaign			-0.078***
			(0.030)
"Hiring-Group" campaign			-0.059***
			(0.021)
Non-Targeted campaign			-0.112***
			(0.022)
Constant	0.155***	0.073	0.102*
	(0.017)	(0.089)	(0.059)
Launch day FE		✓	✓
Standard errors clustering Level	Requester	Requester	Requester
R^2	0.01	0.25	0.33
Number of observations	62,944	62,944	62,944

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients estimates from Equation 5 where the dependent variable is the log payment per task. Standards error clustered at employer level are in parentheses. We add fixed effects for the launch date in columns (2) and (3), and we progressively add controls at the employer and requester levels.

The results in Table 15 indicate that AI campaigns offer higher pay only for data annotation tasks and when controls are not included, column (1). However, when in-

31. We do not find significant difference in the design dimensions explored for campaign we identified having some "hidden AI content". This result possibly resonate with the firm's willingness to "hide" the AI related content of such tasks behind other labels to obtain a more representative sample and do attract only "experienced" worker in data work.

corporating other campaign features, such as the expected time to execute a task and requesters' characteristics, as in column (2) and (3), the wage difference disappears. Indeed, as shown in Table 13, AI tasks, on average, take almost 9 minutes longer, and task duration positively correlates with wage setting. We also observe a substitution pattern between use of "auto-rated" option and task payment. "Auto-rated" is a payment service, and probably to compensate for its cost, requesters set lower wages. The negative correlation resonates with a substitution between ex ante incentives and ex-post moderation. Since tasks are verified more accurately by the platform and poor quality is easily detected, there is less importance in stimulating effort through incentives. "Hiring groups" and, surprisingly, also non-targeted tasks associate with higher piece rates.

2.4.5 Regression Analysis on Ex-Post Moderation

In this Sub-Section we look at how requesters behave after a task i in campaign j is completed and submitted by the worker. Precisely, we investigate differences in rejection of tasks between AI related tasks and other tasks. In order to do so, we estimate the following model:

$$Rej_{i,j,e,t} = \beta_1 Annotate_j + \beta_2 Generate_j + \beta_3 Hidden_j + \sigma_1' X_j' + \sigma_2' Z_{e,t}' + \sigma_3' W_{i,t}' + \delta_i + \phi_t' + \epsilon_{i,j,e,t} \tag{6}$$

Where $Rej_{i,j,t}$ is a binary taking value 1 if a task in campaign j launched by requester e and executed by worker i at time t is rejected and therefore not paid. It takes value 0 if the task is validated. We move the analysis at task level so that we can take into account also the characteristics of the worker who executed the task. In addition to Equation 5, vector ϕ' also includes the execution day fixed effect ($\chi_{k,t}$). While it is not feasible to fully interpret rejection as investigation effort by the requester, considering that it is indeed the equilibrium outcome of investigation but also effort of the worker, we include worker fixed effect (δ_i) to control for time-invariant characteristics of the contributors and various worker time-variant characteristics ($W_{i,t}^{'}$) in order to approach the interpretation

of rejection, as close as possible, to the investigation effort by the requester.

Table 16: Coefficient estimates from Equation 6 for quality investigation

	(1) Task is Rejected	(2) Task is Rejected	(3) Task is Rejected
Data annotation and transcription ($Annotate_j$=1)	0.057***	0.054***	0.046***
	(0.013)	(0.011)	(0.013)
Data Collection/Mining/Extraction/AI training ($Generate_j$=1)	0.018	-0.001	-0.017
	(0.013)	(0.015)	(0.013)
"Hidden" AI ($Generate_j$=1)	0.003	-0.010	-0.000
	(0.011)	(0.008)	(0.006)
Log Payment per Task ($, USD)		0.097***	0.110***
		(0.030)	(0.026)
Log Expected task duration (#minutes to finish)		0.008*	0.005
		(0.004)	(0.003)
"Auto-Rated" campaign		-0.036**	0.017
		(0.017)	(0.012)
Log Campaign size (#available positions)		-0.014***	-0.002
		(0.003)	(0.001)
Log Length instructions (#words)		-0.003	-0.002
		(0.003)	(0.004)
Log Tot. campaigns launched requester		-0.011***	-0.006**
		(0.002)	(0.003)
Log Days since requester's sign-up		0.004*	0.000
		(0.002)	(0.002)
Log Success rate worker		-0.264***	-0.213***
		(0.006)	(0.007)
Log Days since worker's sign-up		0.001	0.002**
		(0.001)	(0.001)
Log Tot. tasks executed by worker		0.002*	0.003***
		(0.001)	(0.001)
Log GDP per capita requester's country (2020)		-0.015***	
		(0.003)	
Constant	0.026***	1.408***	1.000***
	(0.005)	(0.032)	(0.039)
Worker FE		✓	✓
Execution Day FE		✓	✓
Launch Day FE		✓	✓
Target zone FE		✓	✓
Requester FE			✓
Clustering Level	Requester	Requester	Requester
R^2	0.001	0.302	0.446
Number of observations	6,392,252	6,392,252	6,392,252

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table reports the coefficients estimates from Equation 6. Standards error clustered at employer level are in parentheses. In column (1) there are neither fixed effect nor controls. We include controls and fixed effects in columns (2) and (3). Column (3) additionally include employer fixed effect. The linear probability model is estimated with an OLS estimator.

Table 16 indicates that employers reject tasks from data annotation campaigns more often compared to other campaigns: the probability of task rejection increased significantly (at 1% level) This suggests the possibility of increased investigation efforts for such tasks, emphasizing the importance for requester of accuracy of task execution to avoid introducing bias in the final applications. Results hold even when controlling for campaign, firm and worker characteristics, as in columns (2) and (3). Additionally we show

that higher wages correlate with more rejection. The negative correlation between the GDP per capita of the country of the employer and rejection may indicate that employers from richer countries reject less because they investigate less. In wealthier countries, where labor costs are higher, employers might be more inclined to accept sub-optimal work rather than invest additional resources in thorough validation processes.

2.5 Conclusions

Our study delves into the utilization of platform-based microtasking for AI related tasks, particularly in data training. Despite potential obstacles, such as the anonymity of user interactions and concerns on data privacy, that could impede the successful use of online platforms for microtasking in collecting human inputs for AI, we highlight the increasing appeal of the scalable and cost-effective option provided by platform labor for outsourcing. Drawing on data from a commercial microtasking platform, our findings reveal a significant surge in demand for AI relate tasks since 2019. This indicates the emergence of a new market on such platforms for AI related tasks that involves data annotation, secondary generation, and collection. Tasks for data collection primarily serves market research purposes. While data annotation often involves identifying emotions and spatial objects for training AI models, workers are also tasked with recording. Consistent with privacy concerns, annotation tasks related to sensitive data are rarely observed.

The study also examines the behavior of requesters in outsourcing data work to achieve quality task execution. Through a regression analysis, the paper identifies that AI related tasks involve a more thorough pre-selection of workers by the requester, specifically through predefined worker groups and geo-targeted demand. Although no significant differences in monetary incentives are observed, the higher rejection probability in AI campaigns compared to other jobs emphasize the increased importance of quality in this domain and a more substantial effort by requesters to scrutinize quality post-execution.

Our findings underscore the significance of quality for requesters in the domain of AI related tasks, offering valuable insights and guidance for new requesters entering the market about the strategies they can employ to ensure quality. Additionally, they inform

platforms about the most frequently used tools by requesters that can be enhanced to become more attractive, especially considering the growing relevance of the crowd-sourced data.

2.6 Appendix

Recent Developments on the Platform

In this section we provide an overview of the general trends observed on microWorkers from 2016 to 2021, updating the analysis conducted by Hirth, Hoßfeld, and Tran-Gia (2011) at the platform's early stage.

Since 2016, microWorkers has experienced an increase in its user base over time, with particularly rapid growth from the third quarter of 2019. The blue line in Figure 26 plots the quarterly evolution of the number of new registrations on the platform.

Figure 26: Quarterly evolution of registration and new active users on microWorkers

Notes: The figure plots the evolution of the number of new workers registered on the platform each quarter (blue line), and the number of workers, registered in that quarter, who executed at least one task after their registration (red line). The blue line only focus on registration by workers and thus disregards users who acted uniquely as requester. Values are normalized with the first quarter of 2016 as the reference period (value normalized at 100). The shadowed area represent the period after January 2020, and points to the beginning of the Covid-19 pandemic.

In the first half of 2020, the platform experienced the largest growth in its history, during the initial stage of the Covid-19 pandemic. As people faced the consequent economic downturn or had more time at home, the platform likely became an attractive alternative source of income or a way to gain some extra money. This is reflected in the number of registrations, which were more than five times larger in the second quarter of 2020 compared to the same period in 2016. The increasing trend in online labor market

registration resonates with the result of Laitenberger et al. (2022) that identifies a causal effect of saturation in the offline labor market on participation in the online labor market. Yet, differently, during the initial months of the Covid-19 pandemic, this trend has not only resulted in more registrations, but also in an increase in the number of active users among new registered, as shown by the red line in Figure 26 which plots the number of workers that registered in each year quarter and executed at least one task on the platform from the sign-up day up to end of April 2021.

Simultaneously, also labor demand on the platform undergone a significant shift over the past few years, as shown in Figure 27. From 2016 to the third quarter of 2018, the number of campaigns launched grew at a constant pace. Yet, from the last months of 2018, demand began to increase more substantially, with an first peak at the beginning of 2020, when the number of launched campaigns was almost two times larger compared to the initial level of our time series. However, in the first half year after Covid-19 pandemic outbreak at the beginning of 2020, the platform experienced a decline in demand. This initial reduction was followed by a slow rise in the latter part of 2020 and a sudden acceleration in the first quarter of 2021. These trends align with the findings of Stephany et al. (2020) of a "down-scaling loss" (i.e. a contraction of the use of online labor market and non-standard work) followed by a "distancing bonus" (i.e. an increase in demand as companies move operations online). During the pandemic, businesses likely first reduced unnecessary expenses. In countries where traditional labor is protected by the state with restrictions on layoffs, it may have been easier for companies to cut contributions from online freelancers instead of regular employees. As businesses adapted to the new situation and developed new needs, such as the increasing importance of e-commerce and digital marketing, they may have relied more heavily on outsourcing labor through online platforms.

The distribution of contributions by location of users follows the well-established Global North - Global South pattern, observed in the literature on platform work (Berg, Rani, et al. 2021), where demand concentrates in wealthier countries of the Global North, and the supply, predominantly in poorer and developing countries largely situated in the

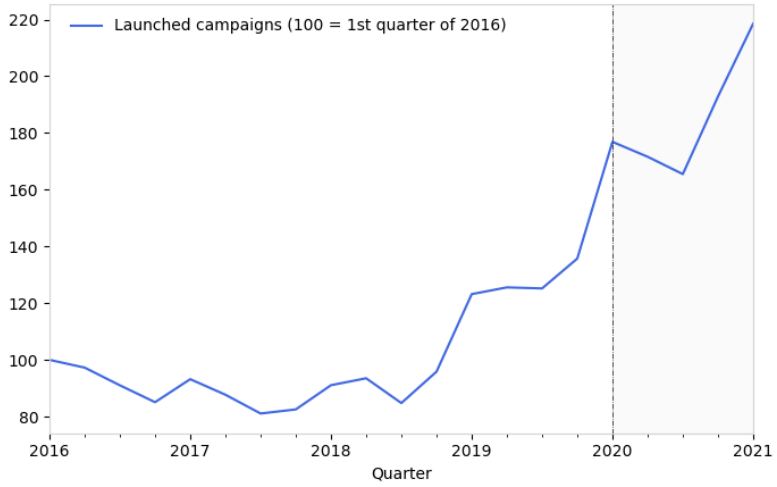


Figure 27: Quarterly evolution of the demand on microWorkers

Notes: The figure plots the evolution of the number of new campaigns launched on the platform each quarter. Values are normalized with the first quarter of 2016 as the reference period (value normalized at 100). The shadowed area represent the period after January 2020, and points to the beginning of the Covid-19 pandemic.

Global South. Despite historically the platform has been predominantly attractive to US workers, this dominance has gradually diminished and workers from other countries have taken a more substantial role in the market. The trend accelerated during Covid-19 pandemic. Figure 28, plots the evolution relative share of contributions by the first five countries of location of workers in the analysed period and shows that the participation of workers from United States declined and transition from being the primary location for workers (representing over 30% of all tasks executed in 2016) to contributing to less than 10% of all finished jobs in the initial months of 2021). Especially after the outbreak of the Covid-19 pandemic the share of tasks executed by workers located in US shrinked significantly replace by growing contributions from English-speaking Asian countries (in particular India) and from Venezuela.

Figure 29 illustrates the evolution of the relative share of campaigns launched by the first five countries where requesters are located in the period of analysis. The figure highlights that demand is predominantly driven by North America (United States and Canada) followed, in a smaller share, by the UK, India and Serbia. In the fist quarter of 2021 demand from English-speaking global North accounted for almost 70% of the total

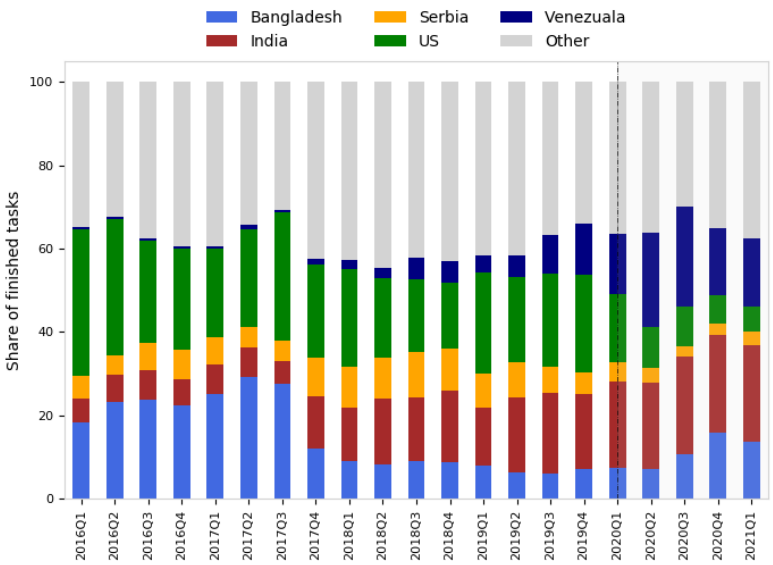


Figure 28: Quarterly share of finished tasks by most common location of workers

Notes: The figure plots, for each quarter, the relative share of tasks finished by workers in the first five locations of the contributors. A sub-set of data annotation tasks are not accounted in the graph, due to lack of complete information on the location of the workers. The shadowed area represent the period after January 2020, and points to the beginning of the Covid-19 crisis.

launch of new campaigns on the platform.

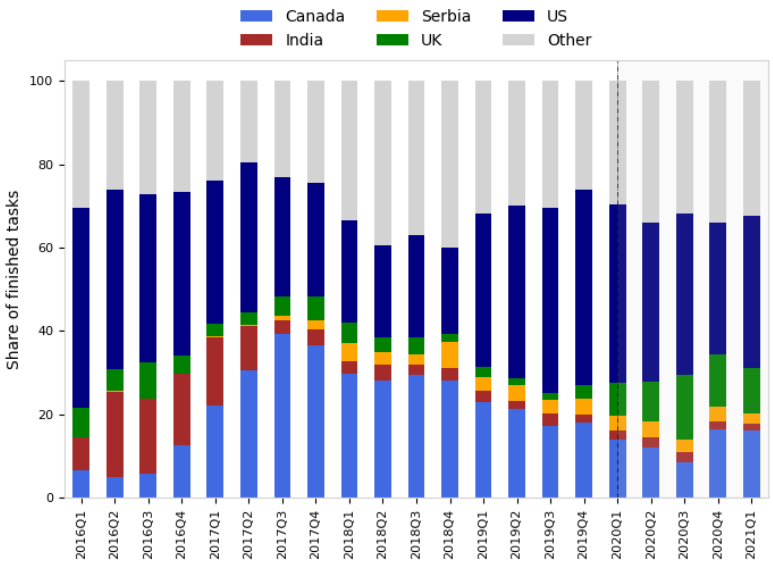


Figure 29: Quarterly share of launched campaigns by most common location of requesters

Notes: The figure plots, for each quarter, the relative share of launched campaigns in the first five locations of requesters. A sub-set of data annotation tasks are not accounted in the graph, due to lack of complete information on the location of the workers. The shadowed area represent the period after January 2020, and points to the beginning of the Covid-19 crisis.

Table 17 returns an idea of the broad scope of the platforms during the period under study. While the majority of tasks have to do with marketing and online promotion (e.g.,

Search Engine Optimization - SEO, offers, and sign-ups), interaction with social media content, there is also demand for data work, such as data annotation, data collection, and transcription. Other tasks involve research related tasks (e.g. answering surveys), Q&A, etc.

Table 17: Campaign frequency at campaign level

Category	Relative Share	Cumulative Share
SEO & Web Traffic	35.099%	35.099%
Offer/Sign up	20.413%	55.512%
Video/Music Sharing Platforms	19.069%	74.581%
Promotion	5.177%	79.758%
Social Media	3.715%	83.473%
Other	2.672%	86.145%
Mobile Applications	2.201%	88.346%
Facebook	1.926%	90.272%
Bookmark a page	1.760%	92.032%
Write an honest review (Service, Product)	1.161%	93.193%
Google (+1)	0.896%	94.089%
Blogging	0.856%	94.944%
Twitter	0.808%	95.752v
Questions, Answers & Comments	0.719%	96.471%
Data Annotation	0.472%	96.943%
Download, Install	0.457%	97.400%
Instagram	0.403%	97.803%
Comment on Other Blogs	0.389%	98.192%
Forums	0.387%	98.580%
Survey/Research Study/Experiment	0.386%	98.965%
Data Collection/Mining/Extraction/AI Training	0.304%	99.269%
Testing	0.254%	99.522%
Leads	0.132%	99.655%
Blog/Website Owners	0.104%	99.759%
Write/Rewrite an Article	0.079%	99.838%
Qualification	0.071%	99.909%
Content Moderation	0.056%	99.965%
Data Transcription	0.017%	99.982%
Content Translation	0.015%	99.997%
Snapchat	0.004%	100.000%

Notes: The table reports the occurrence of campaigns launched between January 2016 and April 2021, categorized by type. The largest category is Search Engine Optimization (SEO) & Web Traffic (35% of the demand). Categories in **bold** are those related to AI.

AI Text Analysis

Keywords: "AI", "Artificial Intelligence","chatbot", "chatbots", "machine learning", "speech", "voice", "voice recognition", "data", "dataset", "datasets", "annotation", "an-

notate", "annotating", "data annotation", "annotated data", "data labeling", "labelling", "label", "labeling data", "data tagging", "tagging data", "face recognition", "image digitization", "object recognition", "object detection", "caption", "captionining", "data transcription", "transcribe", "transcription", "transcribing", "image recognition", "video analysis", "scene understanding", "data classification", "classify", "classification", "classifying", "classification of data", "data categorization", "categorize", "categories", "categorization of data", "categorying", "data preprocessing", "preprocessing of data", "data cleaning", "cleaning of data", "data clean", "data validation", "validation of data", "image processing", "image data", "image feature", "data verification", "verification of data", "data synthesis", "synthesis of data", "data aggregation", "aggregation of data", "sentiment analysis", "sentiment-analysis", "data analytics", "speech recognition", "ai preparation", "data generation", "generate data", "generating data", "image acquisition", "take picture", "take pictures", "take photo", "take photos", "record", "recording", "taking picture", "taking pictures", "camera", "video-camera", "videocamera", "record video", "record videos", "recording video", "recording videos", "recording audio", "record audio", "eye-contact", "mp3", "mp4", "resolution", "jpeg", ".jpeg", ".mp3", ".mp4", "png", ".png", "csv", ".csv", "xlsx", ".xlsx", "video-based surveillance", "data entry", "entry data", "data generation", "generate data", "create data", "generating data", "excel", "row", "column", "cell", "cells", "rows", "columns", "data mining", "ask question", "training alghorithm", "data analytics", "big data", "recognition technology", "language technology", "adaptive learning", "network data", "pattern recognition", "deep-learning", "ML", "machine learning", "feature extraction", "natural language processing", "language processing", "supervised-learning", "unsupervised learning", "un-supervised learning", "semantic search", "computer vision", "visual search", "learning algorithm", "training algorithm", "learning algorithm", "text mining", "Text Mining", "support vector machine", "voice recognition", "voice recognition bioinformatics", "speech processing", "algorithm", "algorithms", "image recognition", "machine learning platform", "data driven model", "automatic recognition".

Figure 30: World cloud of most frequent AI related vocabulary in job's title and description

Notes: The wordcloud show that among the most frequent keywords on our list detected in"hidden" AI campaigns, there is the word "data", words related to data generation such as "record"," camera", "resolution" and to data annotation, such as for example: "label", "categories", "row", "cell", "column".

Additional Results on the Type of AI Campaigns

Table 18: Frequency in different categories of campaign detected as AI related from text analysis but not labelled in a data training campaign

Category	Relative Share	Cumulative Share
SEO & Web Traffic	57.8%	57.8%
Video/Music Sharing Platforms	17.6%	75.4%
Offer/Sign up	16.6%	92.0%
Promotion	2.3%	94.3%
Other	1.2%	95.5%
Write an honest review (Service, Product)	0.8%	96.3%
Mobile Applications	0.7%	97.0%
Social Media	0.6%	97.6%
Qualification	0.5%	98.1%
Forums	0.4%	98.5%
Bookmark a page	0.3%	98.8%
Testing	0.2%	99.0%
Survey/Research Study/Experiment	0.2%	99.2%
Facebook	0.2%	99.4%
Questions, Answers & Comments	0.1%	99.5%
Download, Install	0.1%	99.6%
Content Moderation	0.1%	99.7%

Notes: The table reports the frequency of "hidden" AI campaigns whose tasks were executed between January 2016 and April 2021 by category. For instance, almost 58% of such campaigns "hide" behind the popular SEO & Web Traffic category.

Table 19: Keywords included in each thematic clusters for the "Data Collection/ Mining/ Extraction/ AI training" campaigns

Cluster of words	*Words Included*
Panel A: Required Actions	
Collecting Data	"Capture", "Collect", "Copy", "Download", "Extract", "Find", "Hear", "Mark", "Mine", "Obtain", "Paste", "Retrieve", "Search", "Screenshot", "Visit"
Detecting and Annotating Data	"3d", "Annotate", "Angle", "Box", "Category", "Click", "Classify", "Compare", "Detect", "Distance", "Edges", "Filter", "Height", "Identify", "Label", "Manipulate", "Match", "Notice", "Position", "Scale", "Select", "Shape", "Similarity", "Size", "Spot", "Square", "Tick", "Uppercut", "Watch", "Width"
Generating Data & Providing Data	"Answer", "Complete", "Enter", "Fill" "Provide", "Question", "Reply", "Share", "Survey", "Typing", "Upload"
Moderating	"Moderat"
Solving	"Captcha", "Riddle", "Solve"
Testing	"Test"
Training	"Train"
Panel B: Industries/ Applications	
AI, Captcha, Sound and Image Recognition	"Captcha", "Cough","Eyes", "Face", "Laugh", "Teeth", "Noise", "Pronounc", "Scream", "Smile", "Voice", "Volume", "Whisper"
Apps	" App ", " Apps ", "Mobile app"
Contacts collections	"Contact", "Email adress", "Extract email"
Fitness	" Arm", "Fitness", "Hip", " Leg ", " Legs ", "Lunges", "Knee", "Foot", "Feet", 'hip', "Stretch", "Tricepts", "Plank", "Punch", "Push", "Rope", "Squat", "Touch"
Grocery, Food, Supermarkets	"Chicken", "Coffee", "Drink", "Egg", "Fish", "Food", "Grocery", "Oat", "Oil", "Tea bags", "Prices", "Steak", "Supermarket", "Tomatoes", "Veal", " "Water"
Health & Covid19	"Covid", "Detox", "Doctor", "Mask", "Specialist", "Health", "Hematologist", "Kidneys", "Oncologist", "Pharma", "Prescription", "Sleep", "Slumber", "Treat"
Human Resources	"Curricul", "Cv", 'Skill"
Maps/Spacial Data	"Aerial", "Airborne", "Arena", "Avenue", "Boulevard", "Building", "Bridge", "City", "Cities", "House", "Map", "Neighborhood","Mountain", "Park ", "Parks ", "Road", "Sidewalk", "Street", "Tree"
Real estate, finance	"Estate","Financial", "Homeowner", "Lease", "Loan", "Property"
Research	"Conference", "Research study", "Study", "University"
Restaurants	"Local", "Restaurant", "Yelp"
Riddles	"Clarification", "Riddle"
Sales, promotions, marketing	" Ad ", " Ads ", "Amazon", "Brand", "Business", "Cart", "Company", "Coupon", "Deal", "Ecommerce","E-commerce", "Groupon", "Industr", "Marketing", 'Product', "Purchase", "Retailer", "Sale", "Shipping", "Shop", "Store", "Purchase", "Wishlist"
Vehicles	"Bike", "Car ",, "Cars ", "Cylinder", "Driv", "License", "Vehicle"
Search and Engine Optimization	"Bing", "Browser", "Chrome", "Firefox", "Google", "Yahoo"
Social Media & Blogs	"Blog", "Facebook", "Pinterest", "Social", "Twitter"
Video Games	"Game", "Play"
Videos, Movies, Books	" Book ", " Books ", "Movie","Recipe", "Tutorial", "Tv", "Youtube"
Weather, Nature	"Moon", "Weather"

Notes: The table reports the words included in each cluster used for the detection of the type of "Data Collection/ Mining/ Extraction/ AI training" jobs launched on microWorkers. We cluster words manually. We only focus on English vocabulary and excluded words occurring only once from the list, as well as auxiliary and modal verbs, adjectives, adverbs, number, blurry words, and connectors. We discard words that could occur in more that one cluster.

Table 20: Keywords included in each thematic clusters for the "Data Annotation" campaigns

Cluster of words	Words Included
Panel A: Required Actions	
Add, Type, Draw	"Add", "Digit", "Draw", "Insert", "Type"
Annotate	"Annotat", "Click", "Label"
Evaluate, Classify, Order	"Classif". "Evaluat", "Order"
Answer	"Answer", "Choice"
Listen, Watch	"Listen","Watch"
Select, Search, Check	"Check", "Search", "Select", "Skim"
Suggest and Promote	"Suggest"
Panel B: Input Format	
Audio	"Audio", "Record", "Voice"
Image	"3d", "Image", "Pic", "Photo", "Points cloud"
Video	"Clip", "Video"
Document	"PDF"
Panel C: Industries/Applications	
Cartoons	"Cartoon", "Mask"
Facial expressions detection	"Eye contact", "Emotion"
Insertions and Promotions	"Cart", "Insert"
Maps Annotation and Aerial Pictures	"Aerial" , "Building", "Tree", "Car ", "Cars " Maps

Notes: The table reports the words included in each cluster used for the detection of the type of "Data Annotation" jobs launched on microWorkers. We cluster manually words. We only focus on English vocabulary and excluded from the list words with occurring only once, auxiliary and modal verbs, adjectives, adverbs, number, blurry words and connectors. We discard words that could occur in more that one cluster.

3 Moral Hazard in Micro-Tasking. Evidence from a Structural Model

This chapter is coauthored with Louis-Daniel Pape (Télécom Paris, Institut Polytechnique de Paris).[32]

3.1 Introduction

Online labor platforms are used to generate data to train artificial intelligence (AI) systems. For example, these platforms provide data annotation in the form of survey answers or image recognition services which are necessary to train machine learning algorithms (Ipeirotis, Provost, and Wang 2010; Tubaro, Casilli, and Coville 2020). However, these platforms have characteristics that challenge the quality of the data produced: workers and firms are anonymous and interact sporadically ("on-the-spot"). This environment is therefore prone to moral hazard as workers could be inclined to do their tasks with low effort: for instance, by speeding up the execution time or by providing random answers. The outsourcing firms then need to monitor the quality of a large mass of very small tasks, leading to a trade-off between monitoring costs and data quality. High measurement error in training data may have adverse consequences for the final AI application the firm is developing.

In this paper we study the quality of data annotated on a leading commercial micro-tasking platform. For the analysis, we obtain a sample of data annotation campaigns from the platform. For these campaign, we observe the tasks with which the workers engaged and how these tasks were rated by employers. We augment this data with characteristics of both employers and workers and their past transaction volume and the users' country of origin. However, empirically identifying quality of executed tasks in our observational data is challenging as we only if a task on the platforms has been

32. This work was supported by the French Research Agency (ANR) under grant ANR-19-CE10-0012 ("HUSH") and benefited from the Innovation and Regulation Chair of Digital Services (IRSN). The authors extend their gratitude to Ulrich Laitenberger and Guillaume Thebaudin. Helpful feedback was received at the seminars at Télécom Paris, the AFREN Summer School 2023, the ORG Seminar at the Ludwig Maximilian University of Munich.

validated (and paid) or rejected by the outsourcing firm. We start by conducting a reduced form analysis and show that that raising wages by $10 cents (USD) lowers the probability of rejection by nearly 15 probability points. However, as rejection confounds both effort provision by the worker and task investigation/monitoring from the firm, we cannot interpret rejection directly as a measure of quality. For this reason, we develop a structural model of equilibrium demand and supply for effort to disentangle both. In this model, a worker decides whether to perform a task with effort considering various factors including her effort costs and the expected wage given the probability of being monitored by the firm and an unobserved idiosyncratic shock. The worker infers the probability of firm investigation from observed signals about the firm's investigation cost. The firm takes monitoring decisions based on her cost of investigation and the wage she can recoup if the task is rejected. The equilibrium is defined by rational expectations so that expectations are fulfilled.

We use the platform proprietary data to estimate the parameters of our structural model to quantify the size of moral hazard. With an inner-loop we solve our model system in equilibrium using fixed point iteration and with an outer-loop we search for optimal parameters. We identify the model through a control function based on Petrin and Train (2010). We find that task rejections underestimate proper task execution. As a back of the envelope computation, we calculate that actual low effort rates can be roughly approximated by multiplying the observed rejection rate by a factor of 2.5. Finally, we use our structural model to simulate counter-factual outcomes resulting from alternative incentive schemes. We find that punishing workers who have rejected task by making them pay the task wage would reduce low effort by a substantial 15 percentage points. We also simulate the partial effect of subsidizing firms' investigation with an additional 20% of the recouped wage. This exercise reveals a smaller increase in the probability of worker's effort compared to the simulation of a full-wage penalty on the worker.

Our paper builds on several strands of literature in digital economics, organizational economics, and applied econometrics. The seminal work of Horton (2010) paved the way for a rich interdisciplinary literature on platform mediated piece-work labor. We

contribute to several streams of this literature. First, we add new evidence to the literature that evaluates the quality of executed tasks on micro-tasking platforms. With this scope, Smith et al. (2016) conduct a comparative analysis of the performance of respondents hired on Amazon Mechanical Turk against those of traditional lab respondents. The findings yielded mixed results regarding respondent integrity and data quality, with variations observed across demographic features. Corrigan-Gibbs et al. (2015) detect instances of cheating behavior within the online workforce through an experiment utilizing Amazon Mechanical Turk. However, their study focused on the impact of providing warnings about the consequences of cheating. We do not rely on an experiment as we take the standpoint that a worker's quality provision critically depends on the reputation of the outsourcing firm. A researcher running an online experiment would be doing so as a firm with no reputation, leading to results which are hard to generalize. Our observational data allow us to include this dimension of sellers' reputation in the analysis. Researchers have also provided evidence of limited monitoring in online labor platform. Peer, Vosgerau, and Acquisti (2014) show indeed that employers tend to excessively approve work results, potentially inflating workers' reputations. While a large bulk of the existing literature has proposed new tools for better monitoring and detection of low quality (Ipeirotis, Provost, and Wang 2010; Hirth, Hoßfeld, and Tran-Gia 2013; Agley et al. 2022 and Rivera et al. 2022), screening methods may prove costly and challenging to implement. In addition, our contribution extends to a new understanding of how online labor platforms has been used in the value chain of AI technologies (Duch-Brown, Estrella, et al. 2022; Tubaro, Casilli, and Coville 2020).

We build on the literature on Principal-Agent problems (Arrow 1965). In particular, we contribute to the stream of the literature exploring how quality responds to financial incentives. While some studies looked into the labor supply elasticity to monetary rewards (Dube et al. 2020; Mason and Watts 2010), the impact of monetary incentives on quality of tasks execution shows divergent results. Mason and Watts (2010) run an experiment on AMT and discover that online workers respond rationally to prices but higher pay rates do not necessarily improve work quality, since they improve workers beliefs on

their own value. Ho et al. (2015) find that performance-based payments on AMT work only if sufficiently high and when the task is "effort-sensitive". Workers have their own subjective beliefs about the quality of work required for acceptance, leading them to view fixed payments as implicitly performance-based. In Shaw, Horton, and Chen (2011) the role of financial support results in more accurate performance of tasks executed on AMT when respondents are also encouraged to think about the responses of their peers. Kingsley, Gray, and Suri (2014) identify a low sensitivity of work quality to prices, framing the online labor market as a monopsony. While a substantial portion of these results are obtain in experimental settings, our study broadens the picture by exploring a large observational dataset about the entire market for task annotation on a digital platform for micro-tasking. Our approach is structural and allows us to model the supply and demand sides of the market in order to take into account the role of workers expectation on employer investigation.

Finally, our method relies on the discrete choice literature (Train 2009). We model a simultaneous game of incomplete information as in Bajari et al. (2010) where the decision of one agent depends on those of her principal. We identify the model using the control function approach of Petrin and Train (2010) and maximize a likelihood function which is similar to bivariate probit with partial observability of Poirier (1980). In this sense, this paper contributes to the growing literature using discrete choice models to study labor market outcomes.

The remaining part of the paper is organized as follows: Section 3.2 introduces the empirical setting, details the functioning of the platform under study and describes the data used for the analysis along with relevant descriptive evidence. In Section 3.3, we define the structural model and identification strategy. Section 3.4 delivers both reduced form measures of the sensitivity of rejection to wages, as well as estimates stemming from our estimated structural model. The section concludes with a validation exercise to test the model fit with the data. Section 3.5 discusses the recovered unobserved effort and investigation and put them in relationship with observables. In section 3.6, we simulate counterfactual incentive schemes to improve the quality of the data generated on the

platform. Section 3.7 concludes.

3.2 The Principal-Agent Problem of Micro-Tasking

3.2.1 The Micro-Tasking Industry

Micro-tasking (also known as "click-work" or "crowd-sourcing") involves outsourcing small and repetitive tasks to a group of workers located across the globe and who are paid on a piece-rate scheme. According to Irani (2015), the existence of micro-tasking was motivated by the need to for humans to classify images, texts, and sounds in order to train artificial intelligence systems. As discussed below in Sub-Section 3.2.2, we will focus our analysis specifically on data annotation tasks. Nonetheless, micro-tasking now covers a wide range of other activities. For example, a micro-tasking platform might also be used by a researcher to run an experiment, clean her data, or for collecting survey answers. It can also be used for promotional activities and collecting information for market research purposes. A firm might be collecting feedback on a product or generate web traffic. In general, hourly wages tend to be low despite heterogeneity across workers with an average hourly wage of $3.3 USD (OECD 2021; ILO 2022). ILO (2022) reports of 46 different micro-tasking platforms suggesting the existence of fierce inter-platform competition for both workers and firms.

Our analysis relies on data from one of such platforms. On the platform, firms launch campaigns, as batch of identical small tasks, usually paid between $0.10 USD and a few dollars. They can optionally choose to target a selected group of workers by country of origin or platform rating. The platform has a "generalist" scope, with large range of outsourced jobs, including data annotation, testing, online promotions, survey completion, participating in research experiments, and even simpler actions like clicking on social media links. Tasks are organized into predefined job categories with different minimum payments according which typology of job is required. The category label has to be chosen by the firm. Unlabeled tasks will be assigned to the "other" group. Table 31 in Appendix 3.8 presents an extensive list of all the category labels available on the plat-

form. Firms have access to pre-defined tasks templates or can choose to redirect workers to another other web pages where the tasks are available and design the task their own from scratch. It should be noted that firms and workers do not directly communicate with each other.

Job offers are listed on the platform interface that can be browsed by workers and filtered using criteria such as most recent and highest paid. The platform generates revenue from both firms and workers. It does so by taking a commission for each transaction and charging a fixed fee to employers when they launch a campaign. The firm makes at take-it-or-leave-it offer in the form a unique price for each campaign.[33] However, the platform sets a price-floor according to the type of task (category), and location of workers targeted. Firms can indeed choose workers by location, from previous collaboration or target the demand to pre-defined (by the platform) lists of workers. Workers can view available tasks that match their location and experience in the interface. [34] Public available information about a campaign displayed are the payment, the success rate of previously performed tasks in the same campaign, the share of tasks in the campaign already executed, the estimated time to finish the task (in minutes, based on employer information) and the time the firm needs to rate the output of the worker.[35]

Workers self-select into a task and execute it. Once tasks are submitted, firms have a maximum of 7 days to monitor the quality of a task. They can validate or reject the work. The worker only receives the payment if the task is validated. After the deadline of 7 days, all task are automatically validated and paid, regardless of the quality of their execution.[36]

The problem of tasks performed with low quality has significant economic consequences. First, this problem is general across the broad industry. It affects other microtasking platforms, as well as, more generally, other outsourcing and freelancing platforms

33. All tasks in a campaign are paid the same amount.

34. In practice, the majority of tasks are typically open to all workers with no filtering.

35. We do not include this variable in our analysis because it is almost set at the platform default maximum of 7 days. We do not observe the actual length of time between when a task was executed and when it was rated.

36. The employer can also decide to pay an extra and delegate the monitoring and/or rating process to the platform based on their guidelines.

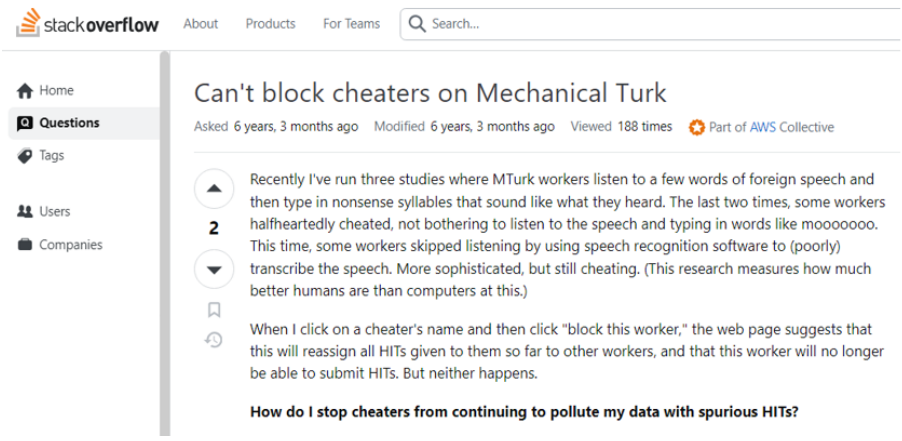


Figure 31: Online forum discussion of quality provision on a micro-tasking platform

Notes: Screenshot from *StackOverflow* (http://archive.today/a190C).

where workers need to work with haste in order to earn a enough income. Second, a task which is improperly done provides no value to the firm. It can even have a negative value if the task is misleading. For example, in Figure 31, a researcher using Amazon Mechanical Turk discusses the problem of "cheaters" on the online forum Stackoverflow. She wants to compare the compare humans to computers in terms of their ability to classify sounds. However, the platform workers provide quasi-random answers in order to save time. If the researcher took these quasi-random answers at face value, she would fail to record human ability, include non-random measurement error in her data, and likely invalidate her analysis. Third, this example reveals the importance for firms in being involved into task monitoring. The latter need to do so to avoid paying for and relying upon poorly done tasks. It also allows the firm to have a reputation which can dissuade the workers from trying to cheat.[37] Finally, this problem of quality directly affects the platform. The need for a reputation represents a barrier to entry for new firms using the platform and the risk of paying for a low quality task limits the value of the platform to potential firms.

We interpret the presence low quality tasks as the result of a Principal-Agent Problem

37. Based on this idea, and in contrast to Dube et al. (2020), we do not rely on experimental evidence (i.e. we do not set up an experiment where we vary the wage to assess its impact on quality) because the response of the worker depends on the reputation of the firm. Without a reputation, we would likely find that workers do not respond to changes in wages because they would presume that an experienced firm is unlikely to investigate and reject their tasks.

(Arrow 1986). That is, we can identify a conflict of interest between the firm (or Principal) and the worker (or Agent). The latter wants to do as many tasks as possible whereas the former needs the quality of the tasks to be good. The agency costs resulting from the deviation from the Principal's interests in the result of asymmetric information: only the worker knows if a task was done with high or low effort (or, equivalently, quality). This forces the principal to investigate (or, equivalently monitor) the agent's work at a cost to the firm (Akerlof 1970). In turn, this creates moral hazard in the form of agents taking a risk over the probability of having their work monitored (or, equivalently, investigated) by the principal (Arrow 1965). The worker ends up supplying too many low quality tasks than is desired by the firm. This interpretation guides the design of the structural model described in Section 3.3 and provides motivation for the different incentive-based counter-factual simulations which we report in Section 3.6.

3.2.2 Data Annotations Tasks for AI Training

Before providing motivating evidence supporting our Principal-Agent based interpretation, we now explain our choice to focus our analysis on a sub-category of tasks called "Data Annotations". These tasks are of particular economic relevance given their role in generating the data which will later be used to train artificial intelligence algorithms and for machine learning. For example, Figure 32 displays the indications given to workers to help them correctly identify trees on an satellite image. This task is simple but can be done poorly, as shown on the right hand side image. One can imagine training a image recognition algorithm based on this data. The algorithm could then be used for aeronautic or military applications. In both cases, accuracy of the final algorithm is key and having accurate data is a necessary condition. It follows that the moral hazard problem, and the associated loss in the quality of the generated data, is a particularly salient problem for data annotation tasks. Moreover these tasks have gained increasing significance on the platform in recent times (Figure 40 in Appendix 3.8). Finally, there is the additional advantage that data annotation tasks tend to be very similar to each other, limiting the risks of unobserved heterogeneity driving the results of our analysis.

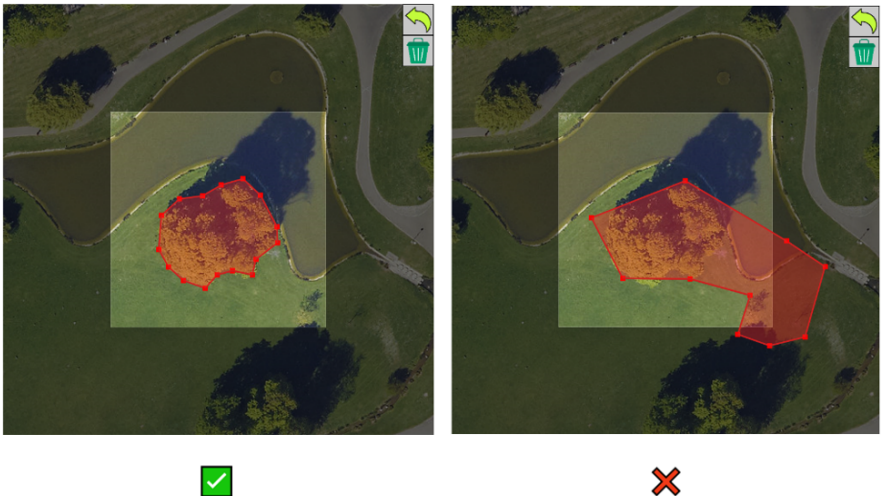


Figure 32: Example of data annotation task

Notes: Data annotation task available on a micro-tasking platform.

Our dataset tracks tasks executed between July 2020 and May 2021 on the platform in the data annotation category on the platform under study.[38] Having data provided directly by a platform is critical to the analysis: data regarding whether a task was accepted or rejected is not public. Moreover, matches between tasks and workers are also unobservable to lone researchers. The data is at the task level and includes information regarding task characteristics (i.e. task descriptions, payment, task outcomes, and expected execution time). We augment this data with characteristics of both firms and workers on their past transaction volume and their country of origin. The final datasets comprises of 20,494 tasks generated by 2,029 workers. These tasks are organized within 1,084 campaigns.

We construct the following set of variables:

- Task is Rejected: Binary variable (dummy) taking a value of 1 if the task faces rejection and 0 if it does not. A rejected task reflects the employer's dissatisfaction with the quality of executed work. In contrast, task validation, mean either a correct task execution or a lack of quality investigation by the firm since unreviewed tasks automatically receive validation after 7 days. The variable mean represents the proportion of tasks that were rejected, a valuable metric for gaining insights into

38. Some data annotation campaigns are excluded from the analysis given the lack of some important information such as the rejection rate.

the overall quality of task completion and the level of firm satisfaction. From our analytical sample, we remove tasks with pending rating status (i.e. the task is not yet validated or rejected). We also exclude tasks who were not rated by the firm but directly by the platform. The structural model at Section 3.3 may be a poor approximation in this case: it is unclear if workers are aware that they will monitored by the platform itself.

- Task Payment: Continuous variable measuring the monetary compensation assigned to the worker if the task is validated is a key metric reflecting the financial remuneration for task completion. The variation of this variable occurs at the campaign level, meaning that within a specific campaign, all tasks receive the same amount of compensation. Firms wage is established at the launch of the campaign and is visible to workers on the platform interface. Since we are interested in the effect of monetary incentives over workers' effort, we drop from the sample tasks with zero wage, such as qualification tests.

- Expected Execution Time: This continuous variable provides insight into the employer's initial assessment of the time required for task completion, serving as a reference for workers and influencing their planning and scheduling during the campaign. It is important to note that this variable does not measure the realized execution time, which is not observed. The expected execution time vary across, but not within, campaigns.

- Campaigns Launched by Firm: We make use of the complete dataset to compute additional statistics concerning the history of users on the platform. We count the cumulative number of campaigns launched by the firm since their entry on the firm.

- Tasks Validated by Firm: We also compute the cumulative number of tasks that the each firm validated (or that were automatically rated after 7 days) since entry on the platform. This figures is updated at task level.

- Tasks Finished by Worker: This continuous variable measure the cumulative sum of

tasks completed by the workers since the moment they registered on the platform.

- Validated Tasks per Worker: Continuous variable counting, task after task, the total number of tasks executed by a worker and validated by the firm.

- GDP per Capita Country of Worker (2020): We complement the platform data with administrative data from the International Monetary Fund (IMF) about the Gross Domestic Product (GDP) per capita in 2020 (expressed in USD) of the country of the worker.[39] It proxy the monetary value of a task given the wealth of the worker's location. To enhance data consistency, we exclude from the sample workers located in countries where this information is not available.

- Outside Option: This continuous variable, with task-day-level variation, counts the number of campaigns available on the platform at a task's execution day.

Table 21: Descriptive statistics of main sample

	N.	Mean	Median	S.D.	Min	Max
Panel A: Task Level Variables						
Task is Rejected (dummy)	20,494	0.10	0	0.29	0	1
Task Payment ($, USD)	20,494	0.16	0.12	0.10	0.02	1
Expected Execution Time (#minutes)	20,494	13.61	10	10.23	2	120
Panel B: Firm-Task Level Variables						
Campaigns Launched by Firm (cumulative sum)	20,494	569	519	267	11	1,493
Tasks Validated by Firm (cumulative sum)	20,494	23,341	24,544	8,141	580	35,830
Panel C: Worker-Task Level Variables						
Tasks finished by Worker (cumulative sum)	20,494	2,334	474	4,947	1	57,948
Validated Tasks per Worker (cumulative sum)	20,494	2,299	462	4,911	0	57,744
GDP per Capita Country of Worker (2020)	20,494	7,949.36	2,270.35	16,244.53	477.38	63,577.34
Outside Option (#available campaigns same day)	20,494	911	807	500	2	2,245

Notes: The table summarizes the distribution of the main variables used in the analysis on our analytical sample of 20,494 observations at task level. Panel A describes the distribution for task-level variables, Panel B for variables at firm-task level and Panel C for variables varying at worker-task level.

Table 21 provides summary statistics for the variables used in our analysis. Panel A shows the distribution of variables at task level. Despite an overall higher rejection rate compared to other task categories (Table 31 in Appendix 3.8), only 10% of data annotation tasks face rejection.[40] Prima facie, this could make one think that employers

39. Worker's country is observed from the self-declared location of the worker.
40. Precisely, 0.095/1.

are often satisfied by the quality of the work. However, at this stage, it is unclear to what degree this number reflects high quality provision from the worker or limited investigation on behalf of the firm. From an econometric standpoint, this rejection rate is sufficiently large to avoid the problem of separation that occurs with binary data subject to very rare occurrences (Albert and Anderson 1984). Payment per task of data annotation tasks range from \$0.02 to \$1 USD. The average wage is \$0.16 USD. The latter is expected from a micro-tasking platform and is similar to the average observed over all categories of the platform (Table 31 in Appendix 3.8). Moreover, it shows the existence of variation in the data which will be exploited for estimating the elasticity of rejection to wages, as well as at the estimation of the structural model of Section 3.3. The estimated completion time varies from 2 minutes to 2 hours. The median indicates that 50% of tasks take less than 10 minutes.

Panel B presents the firm-task level variables of interest. Firms have an average of 569 different campaigns launched. This means that firms are experienced and understand well the platform and are likely to have formed long term beliefs about how to use it profitably. They demand a large amount of tasks: the average of 23,341 tasks validated per firm suggest that demand is high. Regarding worker statistics (Panel C), on average, workers engage in a median number of 474 tasks (462 rated as successful), but there is large variation across workers. There are notable differences in the profiles of the employers and employees in terms of those who have rejected tasks and those who do not. To see this, consider Table 22 which highlights the differences in mean characteristics of the economic agents who engage in tasks which end up rejected versus those with accepted tasks. First, at the task level (Panel A), rejected tasks are, on average, around 35% cheaper than validated tasks and are expected to last 3.5 minutes less. So, rejected tasks tend to be smaller tasks. Second, firms (Panel B) with rejected tasks are less experienced: they have launched fewer campaigns and have validated fewer tasks overall. Finally, workers (Panel C) who have their task rejected are also, on average, less experienced. The workers with tasks rejected have, on average, 1,340 tasks finished fewer than the workers with accepted tasks (i.e. half the experience). Through the structural model

in Section 3.3, we aim to provide a mechanism which can explain why low experienced workers and firms end up with higher rejection rates over tasks which short and less well paid. We will show that the principal agent problem aforementioned, along with varying costs in providing effort and investigation, can explain these patterns.

Table 22: Covariates distribution conditional on task's rejection status

	Task is Rejected		Task is Validated		Difference	
	Mean	S.D.	Mean	S.D	Δ	p-Value
Panel A: Task Level Variables						
Task Payment ($, USD)	0.11	0.03	0.17	0.10	0.06	0.00
Expected Execution Time (minutes)	10.42	5.39	13.95	10.55	3.53	0.00
Panel B: Firm Level Variables						
Campaigns Launched by Firm (cumulative sum)	473.84	135.60	579.47	275.51	105.64	0.00
Tasks Validated by Employer (cumulative sum)	21,581.66	6,683.70	23,526.49	8,258.34	1,944.84	0.00
Panel C: Worker Level Variables						
Tasks Finished by Worker (cumulative sum)	1,121.09	2,995.89	2,461.38	5,092.90	1,340.28	0.00
Successful Tasks by Worker (cumulative sum)	1,089.72	2,954.80	2,426.40	5,056.53	1,336.68	0.00
Ln. GDP Country of Worker (2020)	7.80	0.45	8.15	1.03	0.35	0.00
Ln Outside Option (#available campaigns same day)	6.48	0.63	6.60	0.84	0.12	0.00
Number of Observations	1,947		18,547		20,494	

Notes: The table summarize the mean and standard deviation of the main variables in the analysis for the rejected tasks (columns (1) and (2)) and for the validated tasks (columns (3) and (4)). Column (5) shows the difference and column (6) the p-Value of the t-Test. Rejected and validated tasks are significantly different in all the reported dimensions.

3.2.3 Descriptive Evidence of Moral Hazard

Having described the principal agent problem of the industry in Section 3.2, we now provide suggestive evidence of moral hazard on our dataset of data annotation tasks. This evidence serves to motivate the structural model presented in Section 3.3 and takes the form of three stylized facts concerning the supply and demand for effort/quality in micro-tasking.

Workers bet on having their work investigated. First, workers appear to take a bet over the probability of having their work checked by their employer. Intuitively, if effort is costly and if one's employer never investigated the quality of the work, there is no incentive to provide effort. In a similar fashion, if an employer always investigates the work of her employees, the employee should always provide high effort in order to be paid for her work. Therefore, it follows that workers should trade-off the benefit of providing

low effort against the risk of having their work investigated by their employer (in which case, they will not be paid).

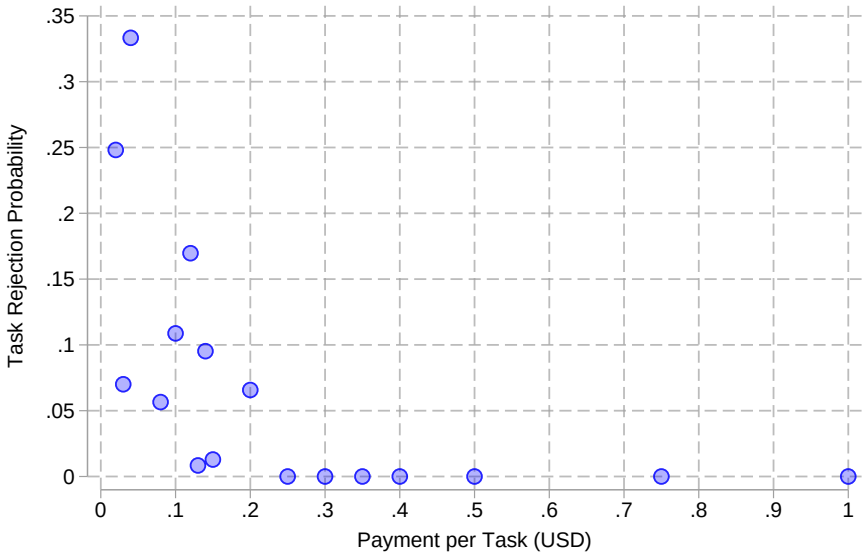


Figure 33: Plot of average probability of rejection given task's payment value

Notes: Binscatter showing the average value of rejection probability for each unique value of payment per task (ln).

As a natural implication of this trade-off, one can expect to find a decreasing rejection rate as wages grow. Indeed, assuming a fixed cost of effort, a higher wage also implies a higher opportunity cost in case one's work is investigated. So, a worker will rather provide higher effort as wages rise. This can be seen in practice by consider the binscatter presented in Figure 33. This figure displays the average rejection rate for each vintile of wage distribution. One observes a monotonic drop from 30% rejection rates for the lowest 10% lest well paid tasks, to a 0% rejection rate for the top 35%. At the same time, the firm has more to loose when wages are high. It follows that they are also more likely to investigate the tasks of the employee. In turn, the worker responds by providing higher effort; creating a retro-active incentive loops.

This evidence motivated our model of effort supply (Equation 13) and demand (Equation 17) presented along with the rest of the our structural model in Section 3.3. In this model, the worker makes a discrete choice between working with high or low effort. Working with high effort comes at a fixed cost, but gives certainty over payment outcomes.

Working with low effort is less costly but may lead to wage loss in case the firm investigates the work and discovers the effort level. Similarly, the firm makes a discrete choice to investigate the task or not. This comes at a cost but with a potential to recuperate the wage. This generates the sort of pattern observed in the figure above.

Risk taking depends on workers' profiles. In addition to rejection depending on wages, there is evidence that the characteristics of the economic agents also affect rejection. For example, Figure 34 illustrates how the past number of tasks executed by the worker (i.e. worker experience) is related to the rejection probability. We observe that rejection is lower for experienced workers and this fall appears to be gradual and economically significant: a worker with no experience is associated with a 20% rejection rate while a worker in the top 5% of the experience distribution has less than a 4% probability of having her task rejected. This may be explained by two potential mechanisms which we explore in the structural model of Section 3.3. One the one hand, there is a cost associated with exerting effort. Most likely, experience provides workers with lower marginal effort costs. As shown in our model of effort supply (Equation 13), this leads workers to to provide higher effort for the same wage, explaining the fall in rejection rate. On the other hand, the firm sets expectations regarding the behavior of the worker. If workers with higher experience have lower effort costs, then they are more likely to provide effort. As a consequence, firms will not have the incentive to investigate them as much. In our model of effort demand (Equation 17), firms investigate less, which also leads to lower rejection rates.

Tasks and Firms characteristics affect tasks' rejection probability. Finally, we now show evidence suggestive that the nature of the task and the characteristics of the employer also have an impact on the probability of a task being rejected. To complement the previous figures, we show this using a linear regression framework which allows us to account for multiple effects at the same time, as well as both firm and worker fixed effects.[41]

41. In Sub-Section 3.4.1, we rely on an instrumental variable strategy to estimate Equation 7 to account for endogeneity resulting from omitted variables.

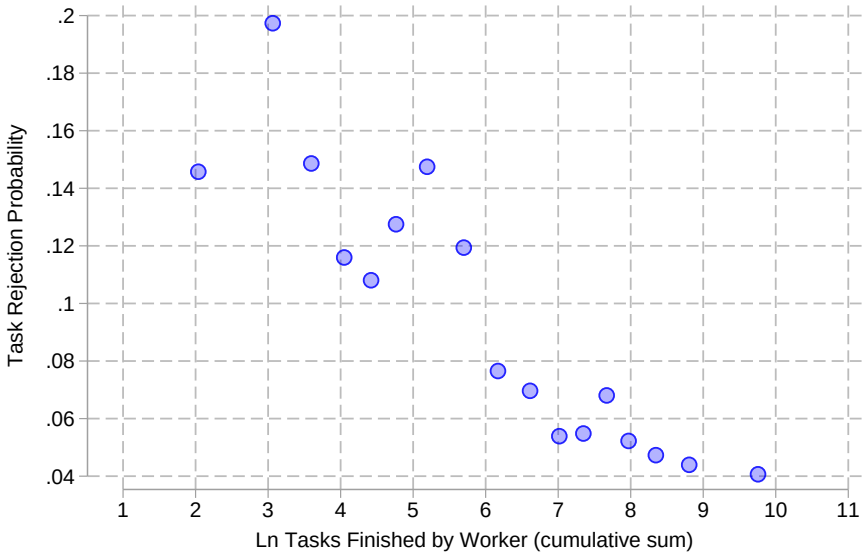


Figure 34: Plot of average probability of rejection given worker's experience

Notes: Binscatter showing the average value of rejection probability for ventile of the distribution of the experience of worker measured by total number of executed tasks (ln).

Our linear model takes the following form:

$$\mathbb{1}\Big(\text{firm } j \text{ rejected worker } i\Big) = \nu + w_j \times \tau + \boldsymbol{C}_j'\boldsymbol{\nu}_j^c + \boldsymbol{C}_i'\boldsymbol{\nu}_i^c + \boldsymbol{T}_j'\boldsymbol{\nu}_j^t + \boldsymbol{\Psi}_i + \boldsymbol{\chi}_j + \xi_{ij}. \quad (7)$$

In this equation, $\mathbb{1}\Big(\text{firm } j \text{ rejected worker } i\Big)$ is a dummy variable equation that takes value one if the task is rejected by the firm. ν is a constant. w_k is the wage per task with associated parameter τ. $\boldsymbol{C}_i$ controls for worker characteristics which can change through time with associated parameters $\boldsymbol{\nu}_i^c$. $\boldsymbol{C}_j$ controls for firm characteristics which can change through time with associated parameters $\boldsymbol{\nu}_j^c$. $\boldsymbol{T}_j$ controls for task characteristics with parameter $\boldsymbol{\nu}_j^t$. Finally, $\boldsymbol{\Psi}_i$ and $\boldsymbol{\chi}_j$ are, respectively, worker and firm fixed effects while ξ_{ij} is an error term with zero mean. We estimate this equation using ordinary least squares with standard errors robust to heteroskedasticity.

Table 23 highlights how the nature of micro-tasks limits the industry's market size. There is a negative and statistically significant correlation between the expected duration of tasks and the probability of rejection. A 10% increase in the expected execution time of a task lowers the probability of rejection by 4 probability points (column (3)). This suggests that task characteristics, such as their time length, can also affect rejection

Table 23: Ordinary Least Square (OLS) estimates

	(1) Rejection Probability	(2) Rejection Probability	(3) Rejection Probability
Task Payment ($, USD)	-0.313***	-0.136***	0.011
	(0.017)	(0.019)	(0.025)
Ln Expected Execution Time (#minutes)	-0.043***	-0.011*	-0.042***
	(0.005)	(0.006)	(0.006)
Ln Campaigns Launched by Firm (cumulative sum)	0.106***	0.007	0.076***
	(0.014)	(0.014)	(0.017)
Ln Successful Tasks by Worker (cumulative sum)	-1.084***	-1.080***	-2.027***
	(0.043)	(0.043)	(0.119)
Ln Tasks Finished by Worker (cumulative sum)	1.087***	1.083***	2.082***
	(0.043)	(0.043)	(0.119)
Ln Tasks Validated by Firm (cumulative sum)	-0.108***	-0.148***	-0.246***
	(0.016)	(0.017)	(0.026)
Ln Outside Option (#available campaigns same day)	-0.013***	-0.007***	-0.011***
	(0.002)	(0.002)	(0.002)
Ln GDP per capita Country Worker (2020)	-0.021***	-0.035***	
	(0.002)	(0.002)	
Firm FE		✓	✓
Worker FE			✓
Standard Errors	Heteroskedastic	Heteroskedastic	Heteroskedastic
R^2	0.135	0.141	0.457
Number of observations	20,494	20,494	20,494

* p<0.1, ** p<0.05, *** p<0.01

Notes: The tables summarizes the coefficients of OLS estimation of a linear probability model. Heteroskedastic standard errors are in parenthesis.

through the supply and demand for effort. As such, longer tasks are perhaps easier or more worthwhile to investigate. This implies a limit to micro-tasking: tasks which are too small result in too low effort rates for them to the worthwhile for employers. This highlight the importance of finding incentive schemes which can help to alleviate the moral hazard problem. This problem concerns not only this platform in particular, but the greater industry of micro-tasks too small to be worthy of the employer's investigation. Given that wages becomes non-significant with worker fixed effects (column (3)), we refrain from further commenting this table in order to address endogeneity concerns in Table 25.

3.3 Structural Model of Effort Provision

We design a game of incomplete information with strategic interactions between the two sides of the platform, workers and firms. The former decide how much effort to provide completing a task, while the latter choose to investigate if the task what done with high effort (or quality). Agents are unable to observe the decision of the other party, requiring them to make decisions in expectations based on public information. Using the

Bayesian Nash Equilibrium as a solution concept, we show that effort and investigation decisions are inter-related. This model yields quasi-closed form predictive probabilities of observing a task being rejected by an employer. These expressions take the form of partially observable bi-variate probability models discussed in the seminal work of Poirier (1980), augmented for strategic interactions. We estimate this model using nested Maximum Likelihood with an inner-loop solving for the Bayesian Nash Equilibrium of each observed match, with an outer-loop searching for the optimal parameters. We demonstrate in-sample fit using auto-rated tasks, as a sanity check and as evidence of internal-validity.

We use this model to disentangle the effort with which a task is done from the frequency with which task are observed to be rejected. While the latter depends on both effort and investigation rates, the former measures directly the quality of tasks. Our estimates reveal that nearly a half of tasks are done without effort and we use our model to understand the underlying mechanisms, as well as to assess how the phenomena relates to the wage scheme. Finally, we use our model to simulate counter-factual outcomes based on a comparative statics exercise consisting in the platform subsidizing firm investigation and rewarding task rejection.

3.3.1 Model Setup

The data consists in matches between workers $i \in \mathcal{I}$ and firms' tasks $j \in \mathcal{J}$. All of the matches belong to a set, denoted by $ij \in \mathcal{M}$. This set is of size N. We summarize all information relevant to the firm by a vector of characteristics $\boldsymbol{C}_j$ and to the particular task by $\boldsymbol{T}_j$. Similarly, workers can be summarized by their characteristics $\boldsymbol{C}_i$. We also observe whether a task is accepted by an employer. We denote this event using a dummy $\mathscr{A}_{ij}$ only equal to zero if it is rejected by the firm. As such, our data $\mathcal{D}_{\mathcal{M}}$ can be written as:

$$\mathcal{D}_{\mathcal{M}} := \left\{ \mathscr{A}_{ij},\ w_j,\ \boldsymbol{C}_j,\ \boldsymbol{C}_i,\ \boldsymbol{T}_j \right\}_{ij=1}^{N}. \tag{8}$$

We model this data through a structural model of task acceptance. The timing of the

game played by the firms and workers is as follows:[42]

- First, firms set up a task along with a take-it-or-leave-it wage proposition w_j. This proposition comes along with an expected tasks duration and other indications for the worker.

- Second, workers can decide to do the task. As discussed more formally below, they either execute the task with high or low effort. The event of a task being done with high effort is denoted by the dummy variable $\mathscr{E}_{ij}$.

- Third, firm j can decide to investigate whether a task is done with high or low effort. We denote by the dummy $\mathscr{I}_{ij}$ which is equal to one only when firm j investigates worker i's task.

- Finally, the worker is paid except if the task is done with low effort and the firm also investigated the task. In this event $\mathscr{A}_{ij} = 0$, the worker is not paid.

According to this model, the observable variable $\mathscr{A}_{ij}$ can be written as the product of low effort provision and firm investigation:

$$\mathscr{A}_{ij} := 1 - \left(1 - \mathscr{E}_{ij}\right) \times \mathscr{I}_{ij} \tag{9}$$

This tells us that the share of tasks which are rejected is an under-estimate of the actual number of tasks which are done with low effort.[43] This under-estimate is proportional to the frequency with which the employer investigates the work of the employee. As we do not observe directly effort and investigation, we must the probability of a task being accepted ($\mathscr{A}_{ij}$). Therefore, we now present the worker and firms decisions to, respectively, provide effort ($\mathscr{E}_{ij}$) and to investigate tasks ($\mathscr{I}_{ij}$). We will then show how it affect the firms' labor demand.

42. Even though steps 2. and 3. occur sequentially, the information used by the agents are both fixed in step 1. As such, their strategies can equivalently be considered as simultaneous.

43. We assume that if the worker exerts effort, the firm will never reject her task. This assumption is supported by the platform requiring firms to provide justification for each task rejected.

3.3.2 Workers' Effort Provision

Worker $i \in \mathcal{I}$ chooses whether to exert effort or not in completing a task. If she exerts low effort ($\mathscr{E}_{ij} = 0$), we assume that the task is done poorly and will be of little value to the firm $j \in \mathcal{J}$. On the other hand, if she exerts effort ($\mathscr{E}_{ij} = 1$), the task is useful to the firm. We take it that providing high effort comes at a normalized cost $\boldsymbol{C}_i'\boldsymbol{\gamma}_i^e + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e$ where $(\boldsymbol{\gamma}_j^e, \boldsymbol{\gamma}_t^e)$ are two vector of parameters which we will estimate. This cost of effort should be interpreted as the opportunity cost of high effort compared to the cost of low effort work (e.g, the additional time and attention required).

Moreover, worker's pay-offs depend on the firm's decision to investigate. That is, the worker providing low effort only receives the wage w_j if the firm does not investigate the task. This contrasts with a worker supplying high effort who will always receive the payment. Since the worker does not know if the firm will investigate her work or not, she builds expectations regarding the firm's behavior. We denote the probability of investigation by i_{ij}. Her mean utility function of exerting effort is:

$$\mathscr{V}_{ij} = \begin{cases} \alpha_i w_j - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e & \text{if } \mathscr{E}_{ij} = 1 \\ \alpha_i w_j \times (1 - i_{ij}) & \text{if } \mathscr{E}_{ij} = 0 \end{cases} \tag{10}$$

where $\alpha_i = \exp(A_i'\boldsymbol{\kappa})$ is the marginal utility to worker i of another dollar which depends on a linear index composed of parameters $\boldsymbol{\kappa}$ and associated variable A_i. As is common in the discrete choice literature Train 2009, we take it these utilities to be subject to match-specific random and independent shocks $\varepsilon_{ij}^{\cdot}$ following a standard Type-1 GEV distribution. As such, we write the expected indirect utility of providing high effort ($\mathscr{E}_{ij} = 1$) as:

$$\mathscr{V}_{ij}^1 = \alpha_i w_j - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e + \varepsilon_{ij}^1 \tag{11}$$

Similarly, the expected indirect utility of providing low effort ($\mathscr{E}_{ij} = 0$) is:

$$\mathscr{V}_{ij}^0 = \alpha_i w_j (1 - i_{ij}) + \varepsilon_{ij}^0 \tag{12}$$

It follows that the probability of effort, denoted by e_{ij}, is given by the logistic c.d.f.:

$$e_{ij} := \Pr\Big(\mathscr{V}_{ij}^{1} > \mathscr{V}_{ij}^{0}\Big) = \Lambda\Big[\alpha_i w_j \times i_{ij} - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^{e} - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^{e}\Big], \tag{13}$$

where $\Lambda[\cdot] := [1 + \exp(-\cdot)]^{-1}$ is the logistic cumulative distribution function. Equation 13 tells us that *ceteris paribus*, increasing the investigation probability i_{ij} results in higher effort provision. This reveals that the quality of a platform is the result of the interaction between workers and firms. Moreover, it shows that higher wages and lower effort costs result in high effort, if there is a non-zero probability of investigation. The lower the cost of exerting effort, the higher the probability of exerting high effort.

3.3.3 Firms' Task Investigation

Firms must decide to investigate the task done by their worker or not. We assume that investigating a task is costly. We model this opportunity cost as $\exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^{i} + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^{i}\right)$, depending on both the characteristics of the firm and of the task. If she investigates a worker's task ($\mathscr{I}_{ij} = 1$), she recoups the wage w_j only when the task is done with low effort ($\mathscr{E}_{ij} = 0$). In the latter case, the firm also avoids using low quality task as measured by a constant s_j which we refer as the scrap value of a task which is no longer usable. When the task is done with high effort, the value of the task for the firm is $\mathscr{Y}_j$. As for the worker, the firm makes decisions based on the expected behavior of the worker. Her mean expected profits are given by:

$$\Pi_{ij} = \begin{cases} e_{ij} \times [\mathscr{Y}_j - w_j] + (1 - e_{ij}) \times s_j - \exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^{i} + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^{i}\right) & \text{if } \mathscr{I}_{ij} = 1 \\ e_{ij} \times \mathscr{Y}_j - w_j & \text{if } \mathscr{I}_{ij} = 0 \end{cases} \tag{14}$$

As for workers, with assume these expected profits to be subject to match-specific independent random error shocks $\eta_{ij} \sim \mathscr{N}(0, \sigma^2)$ which follow a normal distribution with firm specific standard deviation σ. As such, the expected profits of investigating ($\mathscr{I}_{ij} = 1$) are given by:

$$\Pi_{ij}^{1} = e_{ij} \times [\mathscr{Y}_j - w_j] + (1 - e_{ij}) \times s_j - \exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^i + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^i\right) + \eta_{ij}^{1} \tag{15}$$

The expected profit from not investigating ($\mathscr{I}_{ij} = 0$) is:

$$\Pi_{ij}^{0} = e_{ij} \times \mathscr{Y}_j - w_j + \eta_{ij}^{0} \tag{16}$$

Given that the difference between two independent normal distributions is also an independent normal, it follows that the probability of investigating is:

$$i_{ij} := \Pr\left(\Pi_{ij}^{1} > \Pi_{ij}^{0}\right) = \Phi\left[\frac{(w_j + s_j) \times (1 - e_{ij}) - \exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^i + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^i\right)}{\sigma}\right] \tag{17}$$

From this equation, we learn that firms are more likely to investigate when the wage and value of removing a bad task are high, as well as when the probability of effort is low.

3.3.4 Equilibrium

We assume that the probability of effort and investigation are in equilibrium with each other. This results in a system of two equations and two unknowns and we show the equilibrium exists and is unique. To see this, consider writing on the left hand side the probability of low effort using the worker's effort probability, and equalizing it with the probability of low effort contained within the probability of effort provision. One obtains,

$$1 - \Lambda\Big[\alpha_i w_j \times i_{ij} - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_i^e\Big] = \frac{\Phi^{-1}\left[i_{ij}\right] \times \sigma + \exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^i + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^i\right)}{w_j + s_j}. \tag{18}$$

The left-hand side of the equation (effort supply) is strictly decreasing and continuous in i_{ij} over $[0, 1]$. The right-hand side (effort demand) is strictly increasing from minus infinity to plus infinity, as well as continuous. So, the two curves must intersect at some point.

3.3.5 Likelihood Function

Based on Equation 9, we write the probability that a task is accepted as,

$$\Pr\Big(\mathscr{A}_{ij}=1\Big|w_j,\boldsymbol{C}_j,\boldsymbol{C}_i,\boldsymbol{T}_j,s_j\Big)=1-(1-e_{ij})\times i_{ij}= \tag{19}$$

$$=1-\Big(1-\Lambda\Big[\alpha_i w_j\times i_{ij}-\boldsymbol{C}_i'\boldsymbol{\gamma}_i^e\Big]\Big)\times\Phi\left[\frac{(w_j+s_j)\times(1-e_{ij})-\exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^i+\boldsymbol{T}_j'\boldsymbol{\gamma}_t^i\right)}{\sigma}\right] \tag{20}$$

subject to equilibrium constraint (Equation 18).

We can use this expression to construct the following likelihood function:

$$\mathscr{L}\Big(\alpha_i,\sigma,\boldsymbol{\gamma}_i^e,\boldsymbol{\gamma}_j^e,\boldsymbol{\gamma}_i^i,\boldsymbol{\gamma}_t^i\Big|\mathscr{A}_{ij},w_j,\boldsymbol{C}_j,\boldsymbol{C}_i,\boldsymbol{T}_j,s_j\Big)= \tag{21}$$

$$=\sum_{ij\in\mathcal{M}}\ln\Big\{\Pr\Big(\mathscr{A}_{ij}=1\Big|w_j,\boldsymbol{C}_j,\boldsymbol{C}_i,\boldsymbol{T}_j,s_j\Big)\Big\}\times\mathscr{A}_{ij}+\ln\Big\{1-\Pr\Big(\mathscr{A}_{ij}=1\Big|w_j,\boldsymbol{C}_j,\boldsymbol{C}_i,\boldsymbol{T}_j,s_j\Big)\Big\}\times(1-\mathscr{A}_{ij})$$

subject to equilibrium constraint (Equation 18) for each ij $\in \mathcal{M}$.

In practice, we maximize the likelihood function using the BFGS algorithm, while supplying analytic gradients. The optimization procedure involves an inner-loop which takes as given the model parameters and finds the fixed-point $\imath_{ij}^*$ for each match ij by Newton iterations. The outer loop then searches for the optimal model parameters. We calculate standard errors based on the negative of the inverted hessian matrix which we numerically approximate.

3.3.6 Identification through Labor Demand

There are potentially unobserved attributes concerning the task which can be correlated with wages. We address this omitted variable problem using the control function approach of Petrin and Train 2010. To do so, we construct an instrument based on the constraints imposed by the platform on labor demand: the same wage is set for all employees. It

follows that the set wage must then be a function of the characteristics of other employees in the labor markets. We then suppose that conditional on wages, these characteristics of other employees orthogonal to the effort provision of the worker.

To see this more formally, consider the firm's wage-setting problem through the profit function $w_j := argmax_w \pi_j(w)$. We assume that when wages are set, the shocks $(\eta_{ij}^0, \eta_{ij}^1)$ are not yet observed by the firm. Then, denoting the probability that worker i accepts to do task j by $\mathscr{S}_{ij}$, we observe that the wage w_j is a function of the characteristics of all potential employees in the labor market $(\mathcal{I})$, $w_j\Big(C_1, C_2, \ldots, C_N\Big)$, from the following profit maximization problem:

$$\Pi_j = \max_w \sum_{i \in I} \mathscr{S}_{ij}(w) \times \Big[i_{ij}\Big(e_{ij}(C_i) \times [\mathscr{Y}_j - w] + (1 - e_{ij}(C_i)) \times s_j - \exp\Big(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^i + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^i\Big)\Big) + (1 - i_{ij})\Big(e_{ij}(C_i) \times \mathscr{Y}_j - w\Big)\Big] \tag{22}$$

We implement this idea using the "leave-one-out" approach. We first construct the average experience of the workers in the same campaign to instrument the wage per task. More formally, the instrument $\mathscr{Z}_{ij}$ is given by the following expression:

$$\mathscr{Z}_{ij} := (\mathcal{M}_j| - 1)^{-1} \sum_{k \in \mathcal{M}_j - \{i\}} C_k \tag{23}$$

where $\mathcal{M}_j$ denotes the set of workers who performed task j, of cardinality $|\mathcal{M}_j|$. We choose worker experience because it is an economically relevant predictor of effort provision (as seen in Figure 34). We then calculate the mean over the set of workers who actually did the task because it the wage is more likely to reflect the characteristics of workers likely to accept the task over the characteristics of workers who are unlikely to do the task. Figure 41 in Appendix 3.8 provides visual evidence of a linear association between the instrument $\mathscr{Z}_{ij}$ and wages w_j. Second, we regress the task wage on all covariates $(\boldsymbol{C}_j, \boldsymbol{C}_i, \boldsymbol{T}_j)$ appearing in the structural model along with the instrument $\mathscr{Z}_{ij}$,

$$w_j = \alpha + \mathscr{Z}_{ij}\beta + \boldsymbol{C}_j'\boldsymbol{\mu}_j^c + \boldsymbol{C}_i'\boldsymbol{\mu}_i^c + \boldsymbol{T}_j'\boldsymbol{\mu}_j^t + \omega_{ij}. \tag{24}$$

From this equation, one obtain the control function $\hat{\omega}_{ij}$. Finally, we include this

residual as an additional covariate in the cost of effort ($\boldsymbol{C}_i$) and investigation ($\boldsymbol{C}_j$). We provide standard error estimates which are valid for statistical testing under the null hypothesis of no wage endogeneity.

3.4 Estimation Results and Model Validation

3.4.1 Control Function Estimation

We first consider the estimation of the control function. Following Equation 24, we estimate a linear model by ordinary least squares where task payment is the dependent variable and report the results in Table 24. We estimate three specifications: they all includes the control variables ($\boldsymbol{C}_j, \boldsymbol{C}_i, \boldsymbol{T}_j$) but the first one is without any fixed effects. Columns (2) and (3) include firm and worker fixed effects. Column (2), for the purposes of parsimony, is our preferred specification and its residuals are used to estimate the structural model in a second estimation step. The main purposes of these specifications is to explore the range of estimates which can be obtained using an alternative set of identifying assumptions. Standard errors are presented in parenthesis: they are robust to heteroskedasticity.

Several observations can be made on the basis of Table 24. First, the greater the instrument is statistically significant at the 1% level of significance. This is necessary to avoid a weak instrument and, in practice, has greatly improved the speed and accuracy with which our optimization procedure converged, Second, we find that the effect of the instrument is positive. Given that more experienced workers are less often rejected by the employer, they also yield data annotations which are, perhaps, done with higher effort and quality. This would increase the marginal value of a task for the employer and provide an incentive to compensate the worker with higher wages. Third, the associated coefficient is economically significant. In our preferred specification (column (2)): a one task increase in the average experience of the workers having accepted a task in the campaign raises wages by \$0.003 USD. Yet, this measure is nonetheless of moderate size. This suggests that firms are not paying attention solely to workforce experience. Indeed,

Table 24: Instrumental variable estimation: first-stage estimates

	(1) Task Payment ($, USD)	(2) Task Payment ($, USD)	(3) Task Payment ($, USD)
Experience Leave-One-Out Instrument ($\mathscr{Z}_{ij}$)	0.006***	0.003***	0.003***
	(0.000)	(0.000)	(0.001)
Ln Expected Execution Time (#minutes)	0.050***	-0.007**	-0.003
	(0.002)	(0.003)	(0.004)
Ln Campaigns Launched by Firm (cumulative sum)	0.035***	0.169***	0.113***
	(0.006)	(0.008)	(0.013)
Ln Successful Tasks by Worker (cumulative sum)	0.060***	0.040***	0.002
	(0.004)	(0.003)	(0.010)
Ln Tasks Finished by Worker (cumulative sum)	-0.055***	-0.038***	-0.008
	(0.004)	(0.004)	(0.010)
Ln Tasks Validated by Firm (cumulative sum)	0.014**	0.063***	0.072***
	(0.007)	(0.005)	(0.007)
Ln Outside Option (#available campaign same day)	0.003***	-0.006***	-0.003***
	(0.001)	(0.001)	(0.001)
Ln GDP per Capita Country of Worker (2020)	0.004***	0.025***	
	(0.001)	(0.001)	
Firm FE		✓	✓
Worker FE			✓
Standard Errors	Heteroskedastic	Heteroskedastic	Heteroskedastic
Number of observations	20,494	20,494	20,494
R^2	0.423	0.563	0.821

* p<0.1, ** p<0.05, *** p<0.01

Notes: The table summarize the first stage coefficients estimates. Standard errors are in parenthesis. $\mathscr{Z}_{ij}$ represents the average "leave-one-out" (excluded the observed task) skills of the workers who contributed the same tasks' campaign.

there is evidence that wages are also affected by firm-specific characteristics. Both the number of tasks validated and of campaigns (a group of tasks) launched by the employer have a positive and statistically significant relationship with wages. For example, a 10% increase in the number of campaigns launched increases the wage by around $0.17 USD. Finally, adding worke fixed effects (column (3)) do not qualitatively change these observations. The R^2 rises suggesting that substantial variation in wages occurs as a result of firm and worker heterogeneity.[44]

3.4.2 Instrumental Variable Estimates

Before delving into the estimate from the structural model, we consider reduced-form estimates of the elasticity of quality to wages. This allows us to assess the validity of the control function (Equation 24), obtain an estimate based on a standard econometric approach which easily permits the inclusion of employer and worker fixed effects, as well as

44. Incorporating worker and fixed effects in the structural model is a promising improvement which remains computationally difficult. In an unreported specification, the authors kept all firms and estimated the structural model without firm fixed effects. The qualitative findings were akin to those reported in Sections 3.4, 3.5 and 3.6.

to obtain a benchmark against which to compare the estimates from the structural model. To this end, we provide the result from the estimation of instrumental variable estimation of Equation 7 using the first-stage estimates based on Equation 24, in Table 25.[45] The specifications as well as the control variables are the same as for the first-stage estimates in Table 24.

Table 25: Instrumental variable estimation: second-stage estimates

	(1) Task is Rejected	(2) Task is Rejected	(3) Task is Rejected
Task Payment ($, USD)	-1.192***	-1.550***	-0.992*
	(0.112)	(0.253)	(0.520)
Ln Expected Execution Time (minutes)	0.004	-0.020**	-0.045***
	(0.008)	(0.009)	(0.008)
Ln Campaigns Launched by Firm (cumulative sum)	0.150***	0.261***	0.197***
	(0.017)	(0.054)	(0.067)
Ln Successful Tasks by Worker (cumulative sum)	-1.026***	-1.019***	-2.026***
	(0.043)	(0.044)	(0.118)
Ln Tasks Finished by Worker (cumulative sum)	1.033***	1.026***	2.077***
	(0.044)	(0.045)	(0.119)
Ln Tasks Validated by Firm (cumulative sum)	-0.111***	-0.069***	-0.180***
	(0.017)	(0.021)	(0.046)
Ln Outside Option (n. campaigns available at execution day)	-0.010***	-0.016***	-0.015***
	(0.002)	(0.003)	(0.003)
Ln GDP per capita Country of Worker (2020)	-0.013***	0.004	
	(0.002)	(0.008)	
Firm FE		✓	✓
Worker FE			✓
Standard Errors	Heteroskedastic	Heteroskedastic	Heteroskedastic
Number of observations	20,494	20,494	20,494

* p<0.1, ** p<0.05, *** p<0.01

Notes: The tables summarizes the coefficients of the second stage of 2SLS estimation. Standard errors are in parenthesis. Table 24 reports coefficient of the estimation first stage and Table 23 the ordinary least squares estimates.

Table 25 highlights three key characteristics of this labor market. First, there is a negative effect of wages on the probability of a task being rejected. This effect is significant in all three specifications (at the 1% level of significance in columns (1) and (2)). It is also economically significant as a $10 cents (USD) increase in the wage lowers the probability of the task being rejected between 15 (column (2)) and 9 (column (3)) probability points. This suggests that employers could lower the probability of rejecting tasks (and perhaps, also improve the quality of their data annotation tasks) by paying workers more. Similarly, the platform may see task rejections fall as a result of design

45. The specifications correspond to those of Table 23 where estimates are obtained by Ordinary Least Squares.

changes increasing the wages of workers. Second, the reversal in sign and change in magnitude of the coefficients in Table 25 relying on instrumental variables, compared to those estimated by ordinary least squares (Table 23), provides credence to interpreting the leave-one-out instrument $\mathscr{Z}_{ij}$ as not only relevant, but also valid. Finally, the sign, magnitude, and statistical significance associated with the number of tasks done and accepted for each worker is evidence of the mechanism described through the structural model of Section 3.3. Indeed, the probability of a task being rejected seems to depend on the average rate with which a task done by a worker is rejected (the ratio of the two aforementioned variables). This finding is compatible with a firm building expectations over the behavior of her employees (based on her track record) and acting in accordance.

The effect of wage is negative and robust to the inclusion of several controls and fixed effects. Coefficient estimates reveals also a significant impact of other variables on the probability of rejection. Rejection probability tends to be lower for workers with a larger track record of validated past tasks, while they increase with the total number of tasks completed by the worker. Although this may initially seem paradoxical, one possible interpretation could be that past success history can serve as a visible, yet sometimes noisy, signal to firm about worker's quality. Firms may expect high quality if the validation rate of a worker is high, leading to limited investigation effort and lower rejection outcome.

3.4.3 Structural Model Estimates and Validation

This section presents the estimates from the structural model presented in Section 3.3. Table 26 displays the estimated parameters based on maximizing the likelihood function of Equation 21, along with the estimated standard errors, t-statistics, and associated p-values. The results are in line with one's expectations. We find that the marginal utility of income does not vary substantially between countries with high and low GDP per capita. Looking at the cost of effort, we find that the control function is highly significant suggesting the need for correcting for potential sources of endogeneity. As expected, tasks which are longer have higher cost of effort. Similarly, considering the

Table 26: Structural parameters estimates

	Coef. Estim.	Std. Err.	t-Score	p-Value
α_i - Marginal Utility of Wage				
κ: Constant	3.835	0.986	3.887	0.000
κ: Ln GDP per capita country worker (2020)	0.082	0.175	0.470	0.638
$\boldsymbol{\gamma}^e$ - Cost of Effort				
γ_t^e: Constant	-9.500	0.999	-9.513	0.000
γ_t^e: Control Function	15.315	1.000	15.315	0.000
γ_t^e: Ln Expected Execution Time (minutes)	5.871	0.992	5.916	0.000
γ_i^e: Ln Successful Tasks by Worker (cumulative sum)	-27.284	0.957	-28.501	0.000
γ_i^e: Ln Tasks finished by Worker (cumulative sum)	27.623	0.956	28.886	0.000
γ_i^e: Outside Option	-0.771	0.941	-0.820	0.412
$\boldsymbol{\gamma}^i$ - Cost of Investigation				
γ_t^i: Constant	-35.103	0.997	-35.208	0.000
γ_t^i: Control Function	-2.938	1.000	-2.938	0.003
γ_t^i: Ln Expected Execution Time (minutes)	11.305	0.980	11.530	0.000
γ_j^i: Ln Tasks Validated by Employer (cumulative sum)	-2.500	0.708	-3.533	0.000
γ_j^i: Ln Campaigns Launched by Employer (cumulative sum)	0.548	0.889	0.616	0.538
γ_j^i: Employer FE	27.015	0.998	27.080	0.000
$\Phi(\cdot)$ - Effort Demand				
s_j: Scrap Value	-0.053	1.029	-0.052	0.959
σ: Standard Deviation of investigation	-2.685	1.000	-2.686	0.007

Notes: The table summarize the estimates of parameters from our structural model on a sample of 20,494 tasks.

ratio of successful tasks to attempted tasks, workers with a higher success rate have a lower effort cost. The outside option (number of available campaigns available on the platform when the task is executed) is not statistically significant. On the firm side, we find that the control function is also statistically significant. Investigation is more costly for tasks which require more time to execute. The more the worker has validated tasks in the past, the lower the cost of an additional validation, as one would expect. Finally, the scrap value

Before proceeding in exploring the economic implications of our estimates, we assess their reliability through an in-sample and out-of-sample validation exercise. For the former, we compare the estimates to key moments in the data. In Table 27, we first display in Panel A the average rejection observed across the whole sample to the one predicted by the estimates. With a 9% probability, both are virtually the same. We then check in Panel B if the model can match the rejection rates of workers according to their experience (in number of tasks done). This fit is once again quite good, although there

Table 27: In-sample validation

	Estimated Rej.		Observed Rej.	
	Mean	S.D.	Mean	S.D.
Panel A: Whole sample				
All observations	0.093	0.013	0.095	0.086
Panel B: By worker's total number of tasks				
First quartile	0.157	0.027	0.148	0.126
Second quartile	0.097	0.011	0.112	0.107
Third quartile	0.064	0.005	0.067	0.062
Fourth quartile	0.056	0.004	0.044	0.042
Panel C: By firm's total number of validated tasks				
First quartile	0.076	0.006	0.066	0.062
Second quartile	0.099	0.016	0.105	0.094
Third quartile	0.068	0.015	0.068	0.063
Fourth quartile	0.000	0.000	0.000	0.002
Number of Observations	20,494		20,494	

Notes: The table compared the estimated value to the observed one at the sample level (Panel A), by quartile of the distribution of the worker's total number of executed tasks (Panel B) and by quartile of the distribution of the total number of validated tasks by the firm (Panel C).

is a slight over-estimation by the model for the second quartile. Similarly, in Panel C, we evaluate the model through the lense of firms' characteristics. The fit is quite good across the distribution of number of validated tasks. This evidence shows that the non-linear likelihood maximization procedure converged to reasonable estimates which fit the data well.

3.5 Recovery of Unobserved Effort and Investigation

In this section, we explore the economic consequences of the estimates presented in Section 3.4. Using them, we are able to recover previously unobserved economic concepts such as the probability of providing effort for the worker, of investigating a task for the firm, and the share of tasks which should be rejected but are not. In particular, we look at how these probabilities relate to each other and how responsive they are to a change in the payment scheme.

Table 28: Summary statistics of economic estimates

		Mean	Std. Err.	Min.	Max.
Panel A: Effort and Investigation					
$1-\hat{e}_{ij}$	Probability of Low Effort	0.537	0.357	0.000	1.000
$\hat{i}_{ij}$	Probability of Investigation	0.231	0.158	0.000	0.790
$(1-\hat{e}_{ij})\times\hat{i}_{ij}$	Probability of Task Rejection	0.093	0.116	0.000	0.788
$(1-\hat{e}_{ij})\times(1-\hat{i}_{ij})$	Probability of Under-detection	0.443	0.347	0.000	1.000
Panel B: Elasticity with Respect to Task Payment					
$\nabla^e_w \times w_j \hat{e}^{-1}_{ij}$	Probability of Effort (Partial)	0.932	1.209	0.000	13.794
$\nabla^e_w \times w_j \hat{e}^{-1}_{ij}$	Probability of Effort (Total)	1.377	2.591	0.000	50.058
$\nabla^i_w \times w_j \hat{i}^{-1}_{ij}$	Probability of Investigation (Partial)	0.892	1.093	0.000	7.935
$\nabla^i_w \times w_j \hat{i}^{-1}_{ij}$	Probability of Investigation (Total)	0.904	1.510	-0.670	7.327

Notes: The table plots the main object of interest, computed staring from the parameter estimates at Table 26.

3.5.1 Measuring Effort and Investigation Rates

Based on these estimates we compute the main object of interest, as summarized in Table 28. Panel A presents the estimated probabilities for our specific sample. We find that over our sample, around 50% are done with low effort. This can be related to relatively low probability of investigation estimated to be at 23%. Based on this, our model predicts a probability of task rejection of 9%, in line with the data. We then compute the probability of under-detection, which is the the difference in the probability of low effort and task rejection. This measures the share of observations which are low effort but not reported as so by the firm. We find that around 45% of tasks are low effort but not rejected, suggesting that rejection rates are a poor indicator for the actual levels of effort. From a distributional point of view, Figure 44 in Appendix 3.8 shows that effort is distributed following a bi-modal distribution: there are masses near zero effort and a smaller mass of tasks always done with effort. Investigation rates are more normally distributed despite a large mass of tasks which will never be investigated. This also leads to a mass of tasks which will never be rejected.[46]

Turning to Panel B of Table 28, we now focus on how the economic agents respond to changes in task payment (w_j). We estimate two forms of elasticities. Partial elasticities take the response of the other party as fixed. For example, it measures the change in probability of effort assuming that the firm does not increase its investigation rates. In

46. Figure 46 in Appendix 3.8 plots the distribution of the recovered costs of effort and investigation estimates.

contrast, the total elasticity allows for both to change at the same time. Surprisingly, we find that both workers and firms have nearly unit elasticity: a 1% change in the wage results, on average, in a 0.93% increase in the probability of effort (partial). The effects are even larger when consider the multiplier effect induced by accounting for the endogenous response of the other party. For example, a 1% increase in wages raises, on average, the probability of investigation 0.9% when accounting for the way this increase in wages will also increase the probability to supply effort. Looking at these estimates from a distribution standpoint, Figure 45 in Appendix 3.8 shows the existence of an exponential distribution for partial elasticities with a long tail. The spread increases when one considers the equilibrium effect captures by total elasticities: there is even sign reversal showing that in some cases, a rise in wages can end up decreasing effort and investigation. These cases remain reasonably rare.

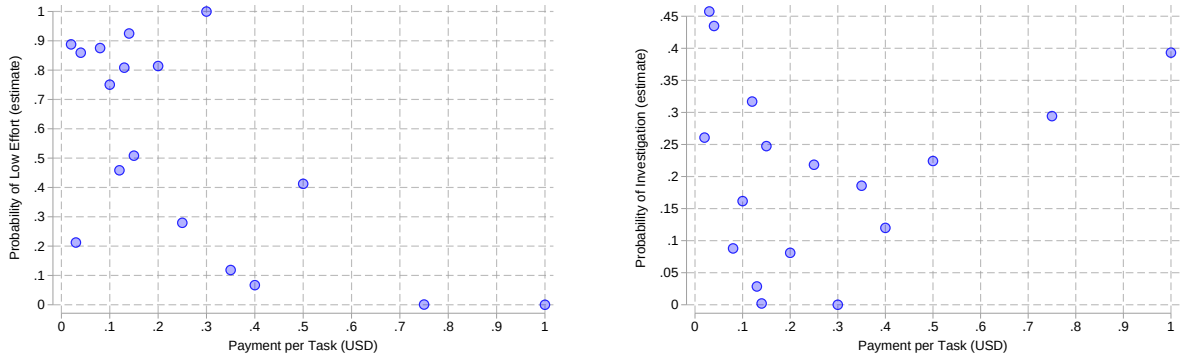
Table 29: Interrelation of rejection, effort, and investigation

	High Rej. Prob. $(1-\hat{e}_{ij}) \times \hat{i}_{ij} \geq 0.5$		Low Rej. Prob. $(1-\hat{e}_{ij}) \times \hat{i}_{ij} < 0.5$		Difference	
	Mean	σ^2	Mean	σ^2	Δ	p-Value
$1-\hat{e}_{ij}$: Prob. of Low Effort	0.98	0.03	0.53	0.36	-0.45	0.00
$\hat{i}_{ij}$: Prob. of Investigation	0.59	0.04	0.23	0.15	-0.36	0.00
Number of Observations	317		20,177		20,494	

Notes: The table summarize the mean and variance of the estimated probability of low effort and investigation according the estimated probability of rejection. The last two columns show the difference among the two groups and highlight the p-Value of a t-t=Test. Both probability of low effort and investigation differs across the two groups.

To better understand how the behavior of workers and firm interact with each other, consider Table 29. It shows that for for the rare cases (317) where there was over 50% probability of the task being rejected, the probability of low effort and investigation are both high, although investigation remains far from 100%. In contrast, for tasks which are unlikely to be rejected (less than 50% probability), the firm and the worker display must lower rates of low effort and investigation. This means that both workers and firms are responding to the incentives faced by the other party, although they do so in a non-proportional way.

3.5.2 Relationship of Observables to Unobservables



(a) Probability of low effort over wage (b) Probability of investigation over wage

Figure 35: Correlation of probability of low effort and investigation estimates with wages

Notes: The figures (binscatter) plot the average value of the estimated probability of low effort (a), and investigation (b) by each ventile of the distribution of the payment per task in USD.

Having recovered unobservable effort and investigation rates through our structural model, we now relate these measures to observable variables. We first focus on how effort and investigation relates to wages. Figure 35 consists in two binscatters of the recovered probabilities against wage. On the left, Sub-Figure (a) shows that the probability of low effort falls a we increase wages. Tasks which are paid \$1 dollar (USD) are found to be done with 0% probability with low effort. However, in Sub-Figure (b), we see that actual investigation rates follow a U-shape. For low wage tasks, workers are providing low effort and this requires the firm to investigate. As wages rise, effort increases allowing the firm reduce its investigation rate. Yet, the price of the most expensive tasks pushes the firm to increase its investigation rate again, although not at the same level as for the lowest paid tasks. This suggests the absence of a monotonic relationship between rejection, investigation, and effort. To see this, one can consider Figure 36. It displays a binscatter of the estimated probability of low effort against the observed rejection probability. We see that the probability that a task is done with low probability and the one that it is rejected co-evolve. However, the gradient is not equal to one: a 10% probability of low effort gives only a 4% rejection rate. So, as a rule of thumb, one should multiply rejection rates by $\times 2.5$ to obtain an approximate low effort rate. This relationship breaks down for tasks done with high probability of low effort. In this case, there is a mass of observations

which are unrelated to the observed rejection rates, reinforcing the idea that rejection rates can be a poor proxy for quality.

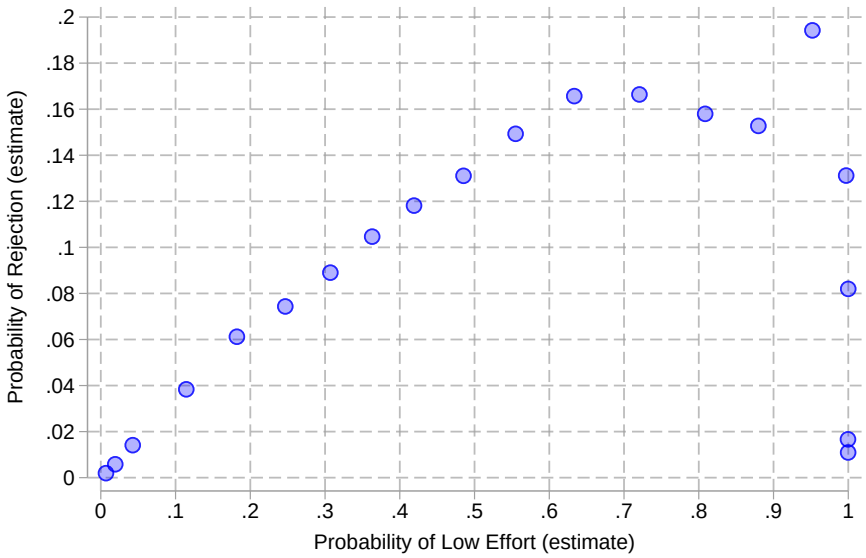


Figure 36: Correlation of probability of low effort estimates with probability of rejection

Notes: Binscatter plotting the average value of the estimated probability of investigation by each ventile of the distribution of the estimated probability of low effort.

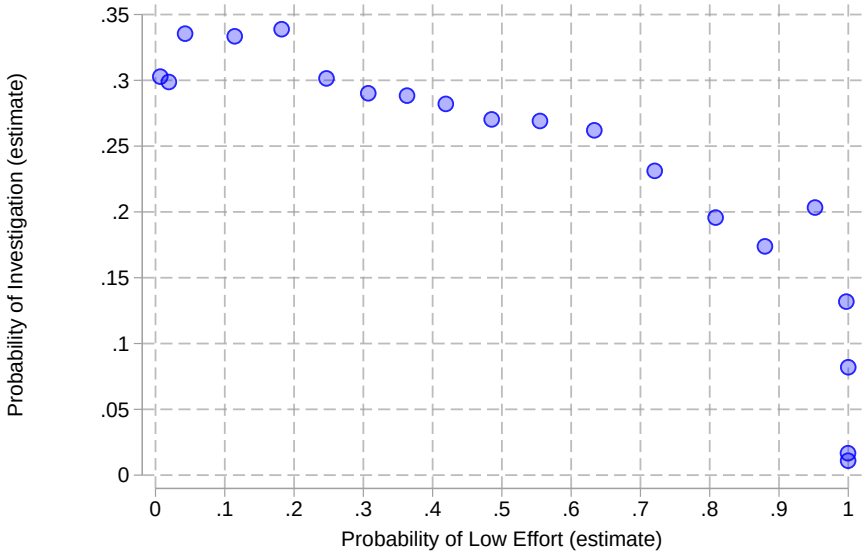


Figure 37: Correlation of probability of low effort estimates with probability of investigation estimates

Notes: Binscatter plotting the average value of the estimated probability of investigation by each ventile of the distribution of the estimated probability of low effort.

Firms may consider their own investigation rate (and their reputation) as a more

meaningful measure of the quality of the tasks. This can be seen from Figure 37 which consists in a binscatter of the investigation probability against the low effort probability. There is a much clearer monotonic relationship: there is a very low chance of having a low quality task when the firm's investigation rate is above 30%. In contrast, below a 15% investigation rate, nearly all tasks will be done with low quality.

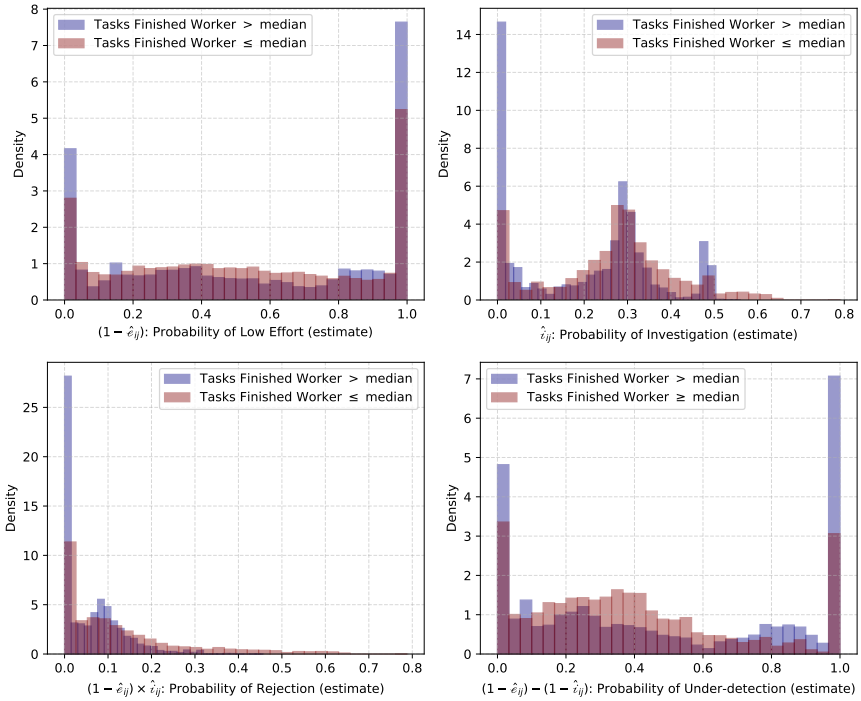


Figure 38: Main estimates distribution conditional on workers' experience

Notes: The above figures plot the distribution of (from top-left to bottom-right corner) low effort probability, investigation probability, validation, and under-detection conditional on the worker's number of executed tasks being above the median (blue) or equal or below it (red).

We now take a distribution approach through Figure 38 which shows that experience (measured as cumulative sum of executed tasks) shifts the cost effort and investigation patterns. There is a bimodal distribution of low effort for workers (top left). Workers with more experience benefit from a greater mass of tasks with very low investigation rates (top right). In turn, the rejection probability is lower at nearly all points across the experience distribution (bottom left). Due to the absence of investigation over tasks finished by experienced workers, there is a large (relative) share of tasks which are very likely to be left undetected (bottom right). The distribution of undetected low effort is more spread for less experienced workers: firms are more unsure about quality and investigate more.

With very experienced workers, the firm never investigates some workers.[47]

3.6 Counterfactual Simulations

The previous section provided evidence that quality was under-estimated by rejection rates. It showed that other indicators, such as a firm's own investigation rate, may work better in practice. Having identified and measured the Principal-Agent problem, we now study the solutions typical of this literature could be applied to our context. To this end, this section relies on the structural model of Section 3.3 to simulate counter-factual policies which could be implemented by the platform. Sub-Section 3.6.1 reports the results from having the platform encourage investigation by providing a subsidy to firms when they reject a task. Sub-Section 3.6.2 focuses instead on providing either a wage subsidy or wage penalty to workers when their task is rejected.[48]

Table 30: Descriptive statistics of simulation results

		Mean	Std. Err.	Min.	Max.
Panel A: Status Quo					
$1-\hat{e}_{ij}$	Probability of Low Effort	0.537	0.357	0.000	1.000
$\hat{i}_{ij}$	Probability of Investigation	0.231	0.158	0.000	0.790
$(1-\hat{e}_{ij})\times\hat{i}_{ij}$	Probability of Task Rejection	0.093	0.116	0.000	0.788
$(1-\hat{e}_{ij})\times(1-\hat{i}_{ij})$	Probability of Under-detection	0.443	0.347	0.000	1.000
Panel B: Subsidy for the Firm (20% of wage)					
$1-\hat{e}_{ij}$	Probability of Low Effort	0.504	0.355	0.000	1.000
$\hat{i}_{ij}$	Probability of Investigation	0.251	0.170	0.000	0.910
$(1-\hat{e}_{ij})\times\hat{i}_{ij}$	Probability of Rejection)	0.097	0.127	0.000	0.890
$(1-\hat{e}_{ij})\times(1-\hat{i}_{ij})$	Probability of Under-detection	0.407	0.346	0.000	1.000
Panel C: Subsidy/Penalty to the Worker (20% of wage)					
$(1-\hat{e}_{ij})$	Probability of Low Effort	0.501	0.366	0.000	1.000
$\hat{i}_{ij}$	Probability of Investigation	0.219	0.152	0.000	0.773
$(1-\hat{e}_{ij})\times\hat{i}_{ij}$:	Probability of Rejection	0.078	0.105	0.000	0.750
$(1-\hat{e}_{ij})\times(1-\hat{i}_{ij})$	Probability of Under-detection	0.423	0.355	0.000	1.000
Panel D: Subsidy/Penalty to the Worker (100% of wage)					
$(1-\hat{e}_{ij})$	Probability of Low Effort	0.403	0.387	0.000	1.000
$\hat{i}_{ij}$	Probability of Investigation	0.190	0.138	0.000	0.645
$(1-\hat{e}_{ij})\times\hat{i}_{ij}$	Probability of Rejection	0.041	0.075	0.000	0.645
$(1-\hat{e}_{ij})\times(1-\hat{i}_{ij})$	Probability of Under-detection	0.362	0.376	0.000	1.000

Notes: The table plots the main object of interest, computed staring from the parameter estimates at Table 26.

47. Tables 42 and 43 in Appendix 3.8 shows the correlation between the estimated probability of low effort and investigation and, respectively, worker's number of executed tasks and firm's number of validated tasks.

48. We run these experiments in partial equilibrium. That is, we do not account for the way in which these platform policies could changes wages, participation into the platform, or the matches between workers and firms. Accounting for these effects would be a valuable direction for developing this research article.

3.6.1 Platform Subsidy for Firms

The previous section of the paper demonstrated that task investigation is necessary for effort provision. We consider a platform subsidy as means to encourage further task investigation. This subsidy takes the form of a payment ω_{ij} paid by the platform to the firm upon a task being rejected. We only set this subsidy to 20% of the wage, ($\omega_{ij} := 20\% \times w_j$) as this policy is costly for the platform. We run this simulation by solving for a new equilibrium investigation rate i_{ij}^S and effort rate e_{ij}^S such that:

$$1 - \Lambda\Big[\alpha_i w_j \times i_{ij}^S - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_i^e\Big] = \frac{\Phi^{-1}\left[i_{ij}^S\right] \times \sigma + \exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^i + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^i\right)}{w_j + \omega_{ij} + s_j}. \tag{25}$$

Results are presented in Panel B of Table 30. On average, the probability of low effort falls by only 3 probability points and the investigation rate increases by 2 probability points. The limited response of the firm to this incentive scheme leaves the rejection change almost unchanged, at 9.7%. As seen in Figure 47 in Appendix 3.8, there are no particular changes in the distribution of investigation and effort probabilities which could salvage this policy. In summary, this policy would be both costly for the platform and have limited impact on investigation, effort, and rejection rates.

3.6.2 Platform Incentive to Workers

As an alternative policy, we now consider two equivalent incentive schemes designed for workers. These incentive schemes consists in the platform providing a subsidy to workers who have a task accepted. This increases their effective wage and encourages effort. However, this policy is costly for the platform given that only a minority of tasks are rejected in practice. In contrast, consider a penalty to worker who would need to pay the platform whenever a task is rejected. This policy would raise revenue for the platform and concern only a minority of workers. From an incentive point of view, these two policies are equivalent because they both increase the payoff-gap between the state of the world where the firm investigates a task and the one where she does not.

To simulate both of these policies, we therefore solve the same fixed-point problem.

Let ζ_{ij} be the subsidy or penalty. We simulate two versions, with a $\zeta_{ij} = 20\% \times w_j$ and $\zeta_{ij} = 100\% \times w_j$ subsidy. The probability of effort (e_{ij}^P) is given equivalently as a penalty or subsidy, as seen from Equation 26 below.

$$
\begin{aligned}
e_{ij}^P(\zeta_{ij}) = & \\
& \frac{\exp\left[\alpha_i(w_{ij}+\zeta_{ij}) - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e\right]}{\exp\left[\alpha_i(w_{ij}+\zeta_{ij}) - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e\right] + \exp\left[(1-i_{ij}^S)\alpha_i(w_{ij}+\zeta_{ij})\right]} \\
& = \frac{\exp\left[\alpha_i w_{ij} - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e\right]}{\exp\left[\alpha_i w_{ij} - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e\right] + \exp\left[(1-i_{ij}^S)\alpha_i w_{ij} - \alpha_i i_{ij}^S \zeta_{ij}\right]} \\
& = \Lambda\Big[\alpha_i\left(w_j+\zeta_{ij}\right) \times i_{ij}^S - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_t^e\Big].
\end{aligned}
\tag{26}
$$

where i_{ij}^S is the investigation probability. It follows that we can simulate these policies by solving the same fixed-point problem:

$$
1 - \Lambda\Big[\alpha_i(w_j+\zeta_{ij}) \times i_{ij}^S - \boldsymbol{C}_i'\boldsymbol{\gamma}_i^e - \boldsymbol{T}_j'\boldsymbol{\gamma}_i^e\Big] = \frac{\Phi^{-1}\left[i_{ij}^S\right] \times \sigma + \exp\left(\boldsymbol{C}_j'\boldsymbol{\gamma}_j^i + \boldsymbol{T}_j'\boldsymbol{\gamma}_t^i\right)}{w_j + s_j}. \tag{27}
$$

The results are displayed in Panels C and D of Table 30. The results from subsidizing the workers by 20% are, on average, the same as for subsidizing the firm. There is slightly less investigation and therefore rejection. However, the probability of low effort is the same (see Panel B). One should note that subsidizing workers for accepted tasks is more expensive for the platform in comparison to subsidising firms for rejecting (an event which occurs rarely).

To go beyond the financial constraints which make subsidies unappealing for a platform, we now discuss a significant penalty set on workers who have a task rejected. Panel D of Table 30 shows that the probability of low effort would drop by an average of 13 probability points. Investigation would also fall by 4 probability points due to the lessened need to investigate well incentivised tasks. This leads to a drop of the rejection rate by nearly 50%. The share of the data which results from low effort and which is not rejected falls by 20%. Figure 39 provides distributional evidence. It shows that the policy

affects the mass of tasks with a very high probability of low effort (top left). This mass is redistributed at the bottom of distribution suggesting that the policy is well targeted. This results in an important mass of tasks which will never be rejected (bottom left) and therefore never detected as stemming from low effort (bottom right).

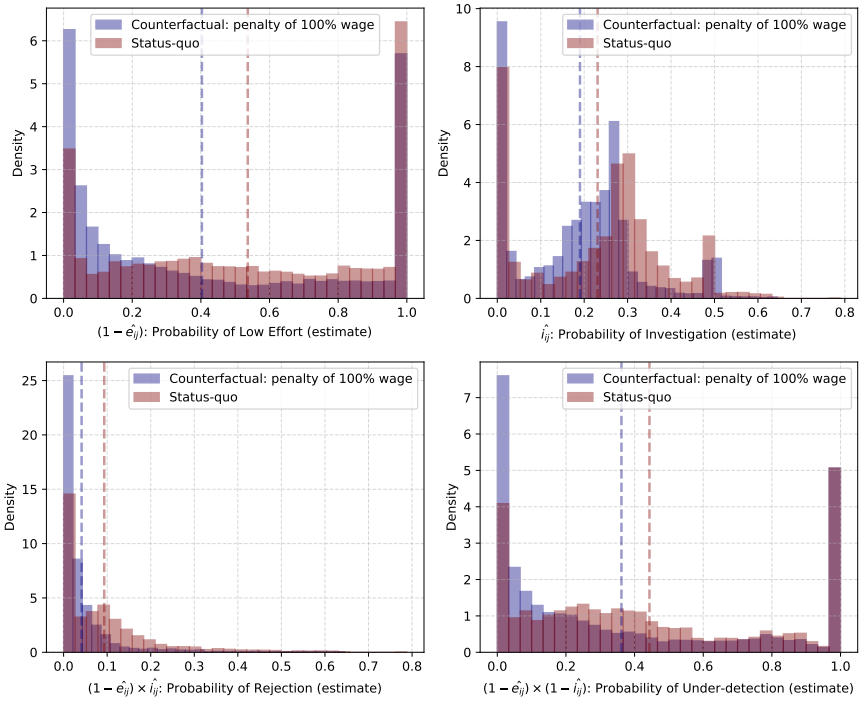


Figure 39: Comparative statics: 100% penalty/subsidy to the Worker

Notes: The above figures plot the distribution of (from top-left to bottom-right corner) low effort probability, investigation probability, validation, and under-detection for the status quo (red) and for the counterfactual simulating a penalty of 100% of the wage to the worker (blue).

This policy is therefore significantly more potent as well as direct source of revenue for the platform. By being cheap to implement, it is possible to raise the salience of the policy. A further refinement which would limit the fall in investigation rates would involve subsidizing the firm who rejected a task with the money taken from the worker who had her task rejected. Although we have assumed that the subsidies and penalties were fixed percentages of the original wage, one could also study how heterogeneity in the incentive system could further improve the overall quality of the work performed on the platform.

3.7 Conclusions

In conclusion, this paper conducts an assessment of the quality and monitoring of data annotation work on a commercial online micro-tasking platform, exploring the effects of monetary incentives. These platforms play a crucial role in tasks related to data annotation, contributing significantly to algorithmic training. However, certain market features, such as piece-rate-limited compensation, anonymity, and limited repeated interactions, may incentivize moral hazard among workers and task fragmentation could constrain investigation efforts by firms.

Quality remains unobserved, while the outcome of rejection and validation decision are observable for each single task in the platform data. However, validation is limited in informativeness when stemming from a lack of investigation by the firm. To disentangle the mechanisms governing rejection, namely the worker effort and the firm investigation decision, the paper adopts a structural approach, modelling the simultaneous demand and supply of effort on the platform.

The model considers the moderating impact of expectations from each platform's side on the other side's choice: the value of wages for the worker is influenced by the expected investigation they will undergo and; similarly firms take into account the expected effort by the worker when deciding if monitoring quality of tasks. The equilibrium outcome, observed as rejection/validation decisions in the data, is derived through fulfilled rational expectations.

Crucially, this research provides platforms with a tangible measure to assess work quality through the observed data of rejection and investigation, offering valuable insights for the platform operator. This includes rules of thumb for evaluating quality based on one own's investigation and rejection rates. Through various counterfactual simulations, the paper demonstrates that monetary incentives can play a role in enhancing task quality and mitigating the risk of introducing undetected biases in the final work applications.

3.8 Appendix

Additional Results

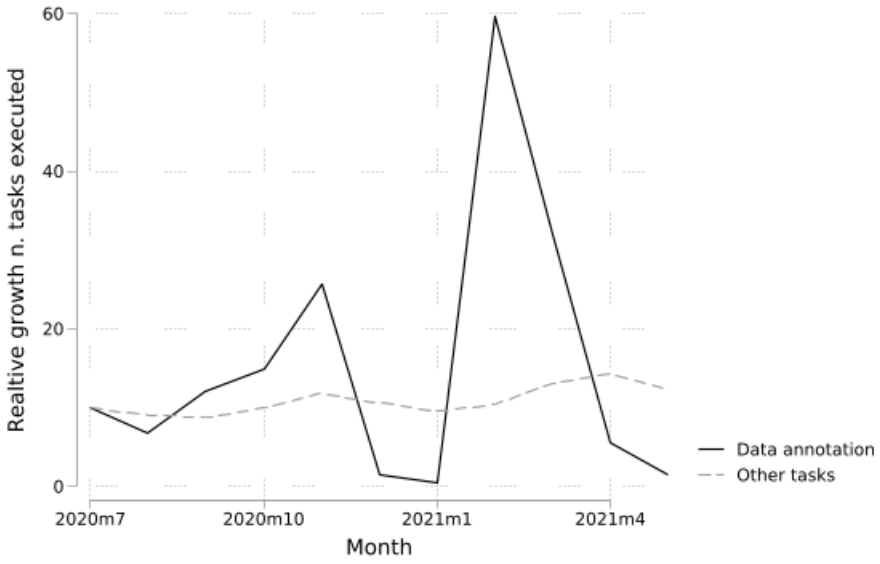


Figure 40: Relative growth number of executed tasks over time on the platform

Notes: The figure compares the relative evolution (compared to initial value at July 2020) of tasks executed in the category "Data Annotation" (solid line) and tasks in other categories (dashed line). The figure illustrates the substantial increase in the number of tasks completed in the data annotation category, showing a notable spike of +500% in February 2021 compared to July 2020. This relative growth is significantly larger compared to the other type of tasks outsourced on the platform in the same period.

Table 31: Frequency, rejection and average wage by task category on the platform

Category name	Frequency	Average Wage	Rejection Rate
Blogging	0.08%	0.44	3%
Content Moderation	0.31%	0.05	4%
Content Translation	0.00%	0.36	9%
Data Annotation	0.24%	0.16	10%
Data Collection/Mining/Extraction/AI Training	1.86%	0.08	3%
Data Transcription	0.00%	2.48	14%
Download, Install	0.24%	0.49	21%
Forums	0.09%	0.15	16%
Leads	0.02%	1.11	45%
Mobile Applications (iPhone & Android)	1.14%	0.43	7%
Offer/Sign up	20.17%	0.18	6%
Other	1.48%	0.17	8%
Promotion	13.03%	0.07	1%
Questions, Answers & Comments	0.37%	0.22	11%
SEO & Web Traffic	24.71%	0.11	1%
Social Media	6.87%	0.15	1%
Survey/Research Study/Experiment	0.50%	0.81	3%
Testing	0.20%	0.44	4%
Video/Music Sharing Platforms	25.29%	0.11	2%
Write an honest review (Service, Product)	3.40%	0.24	4%
Write/Rewrite an Article	0.01%	0.87	7%

Notes: The table reports the frequency of tasks executed on the platform from July 2020 to April 2021, the average wage and the rejection rate (share of rejected tasks) in each category. "Data Annotation" category is in **bold** as it is the focus of this study.

Table 32: Descriptive statistics analytical sample (Ln variables)

	N.	Mean	S.D.	Min	Max
Task is Rejected (dummy)	20494.00	0.10	0.29	0.00	1.00
Ln Task Payment ($, USD)	20494.00	0.15	0.08	0.02	0.69
Ln Expected Execution Time (#minutes)	20494.00	2.55	0.45	1.10	4.80
Ln Tasks Validated by Firm (cumulative sum)	20494.00	9.89	0.81	6.36	10.49
Ln Tasks Finished by Worker (cumulative sum)	20494.00	6.02	2.13	0.69	10.97
Ln Validated Tasks per Worker (cumulative sum)	20494.00	5.97	2.17	0.00	10.96
Ln GDP per capita Country of Worker (2020)	20494.00	8.11	0.99	6.17	11.06
Ln Outside Option (# available campaigns same day)	20494.00	6.59	0.82	1.10	7.72

Note: This table summarize the distribution of the main variables in log used in the analysis on our analytical sample of 20,494 observations at task level.

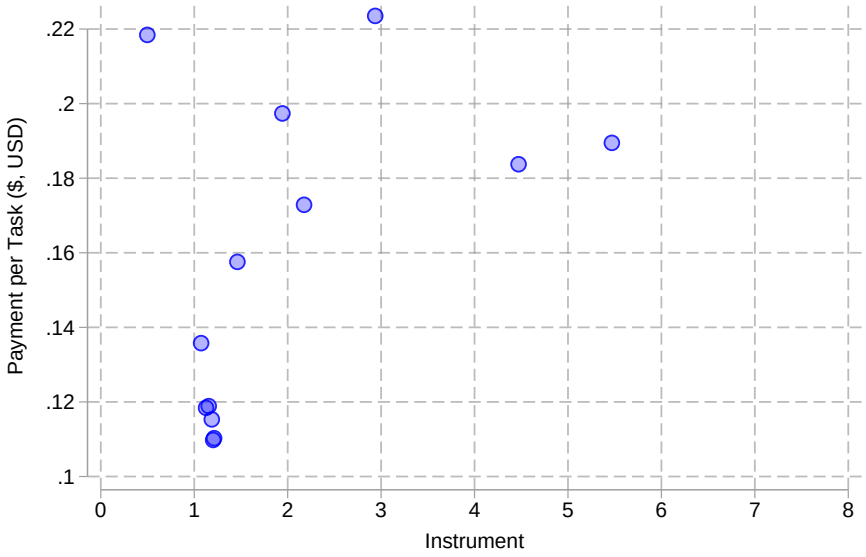
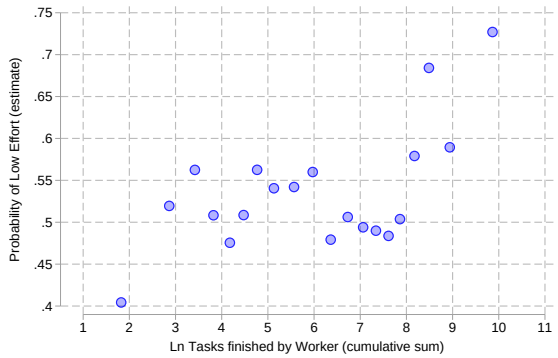


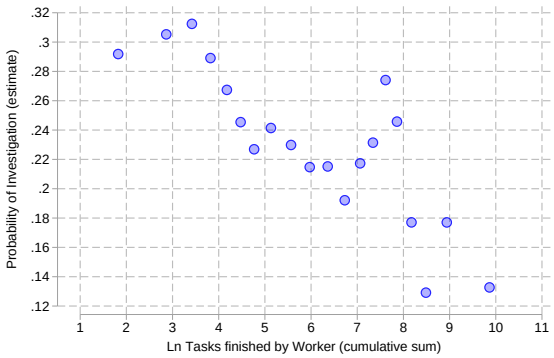
Figure 41: First-Stage: wage against instrument $\mathscr{Z}_{ij}$

Notes: Binscatter plotting the average value of payment per task for each ventile of the distribution of the instrument $\mathscr{Z}_{ij}$.

Additional results structural estimation



(a) Probability of low effort over worker's n. tasks

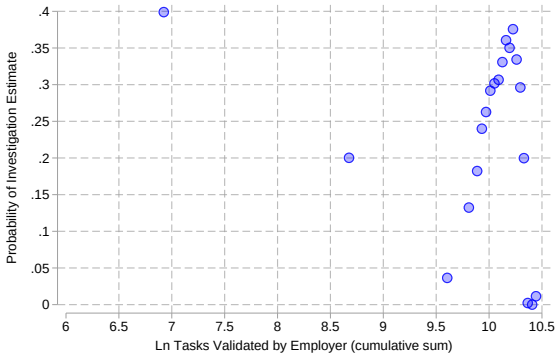


(b) Probability of investigation over worker's n. tasks

Figure 42: Correlation of estimates low effort and investigation probabilities with worker's experience

Notes: This binscatter plots the average value of the estimated probability of low effort (a) and investigation (b) by each ventile of the distribution of the worker's number of tasks executed.

(a) Probability of low effort over firm's validated tasks



(b) Probability of investigation over firm's validated tasks

Figure 43: Estimated probability of low effort and investigation over firm's total number of validated tasks

Notes: This binscatter plots the average value of the estimated probability of low effort (a) and investigation (b) by each ventile of the distribution of the firm's number of tasks validated.

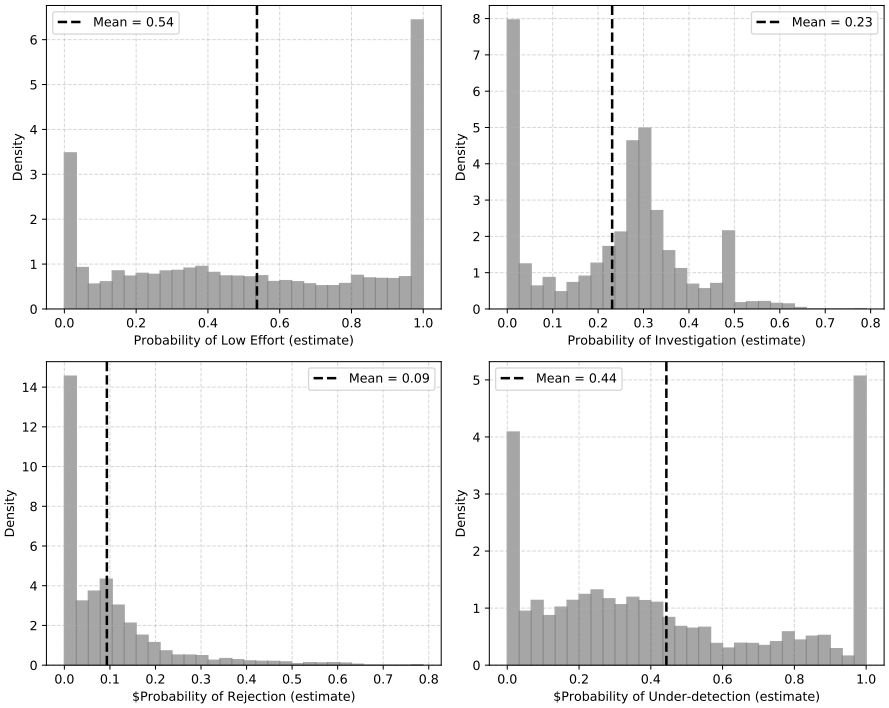


Figure 44: Distribution of main estimates

Notes: The figures illustrate the distribution of the main object of interests from our structural estimation (probability of low effort, probability of investigation, rejection probability and under-detection probability. Vertical dashed lines mark the mean variable.

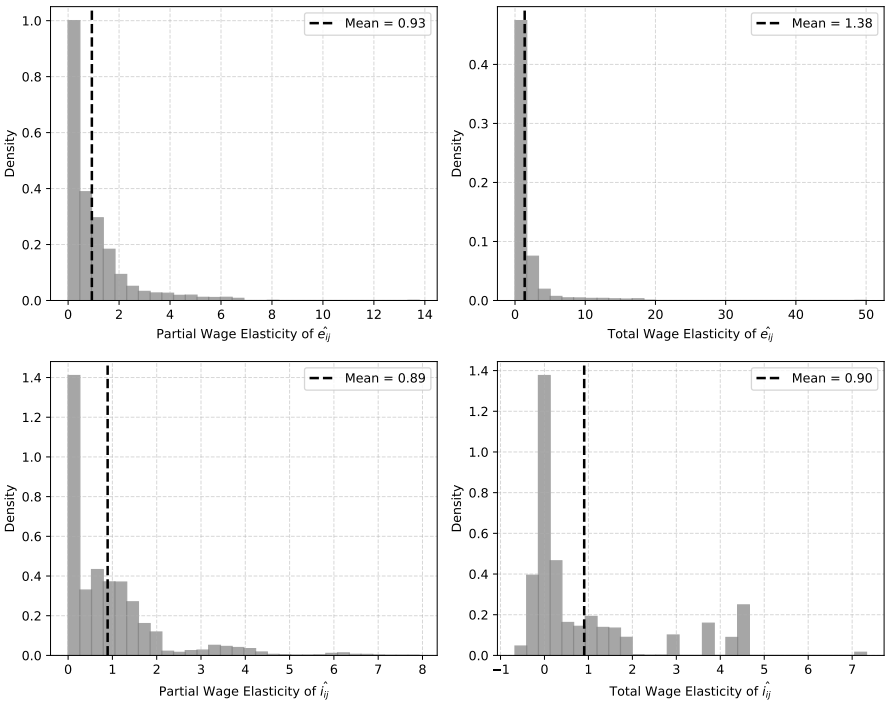


Figure 45: Distribution of estimated effort and investigation elasticities

Notes: The figures illustrate the distribution of the elasticities obtained from our structural estimation. Vertical dashed lines mark the mean variable.

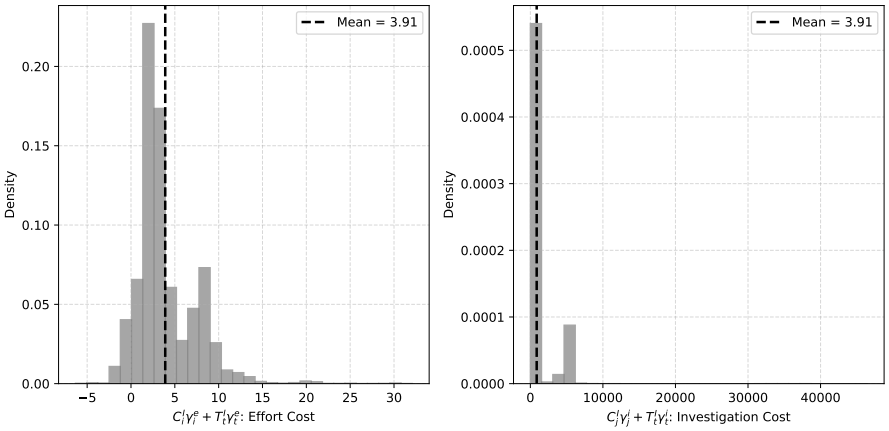


Figure 46: Distribution of estimated cost of effort and investigation

Notes: The figures illustrate the distribution of the estimated cost of worker's effort and firm's investigation. Vertical dashed line marks the mean variable.

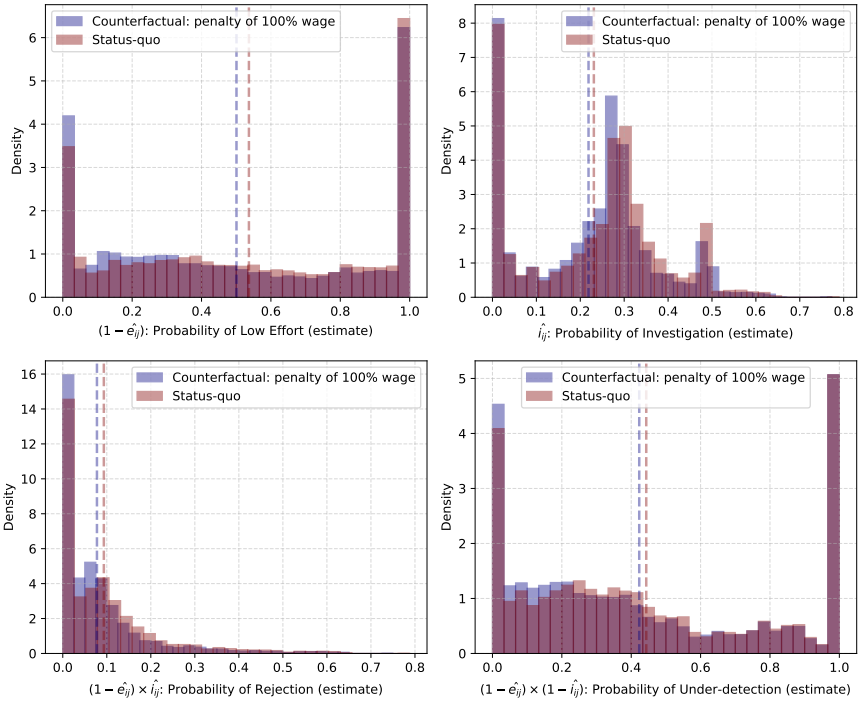


Figure 47: Comparative statics: 20% rejection subsidy to firms

Notes: The above figures plot the distribution of (from top-left to bottom-right corner) low effort probability, investigation probability, validation, and under-detection for the status quo (red) and for the counterfactual simulating a rejection subsidy for the firm of 20% of the wage (blue).

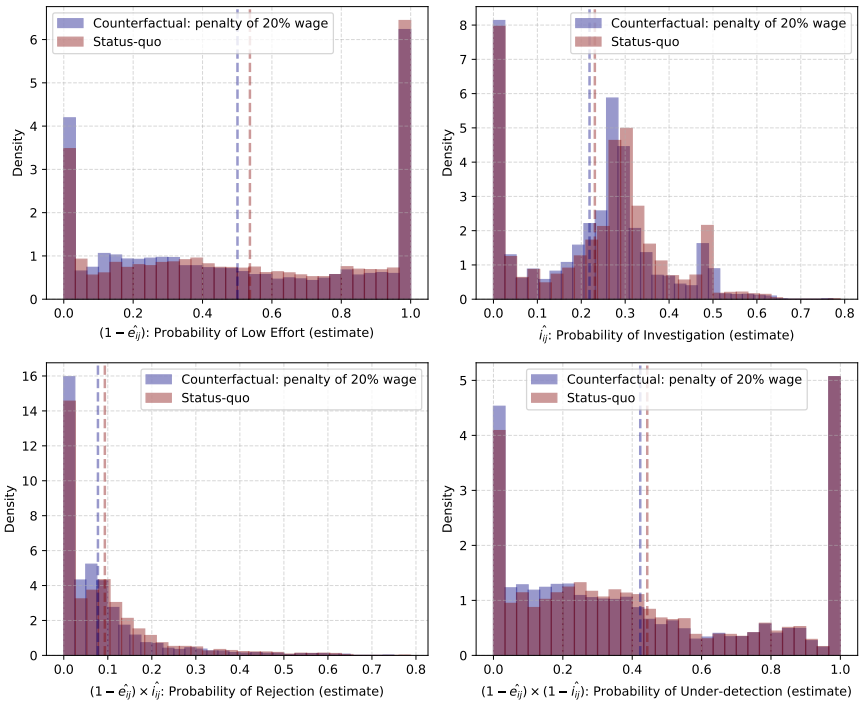


Figure 48: Comparative statics: 20% rejection penalty to the worker

Notes: The above figures plot the distribution of (from top-left to bottom-right corner) low effort probability, investigation probability, validation, and under-detection for the status quo (red) and for the counterfactual simulating a penalty of 20% of the wage to the worker (blue).

4 Conclusions

This thesis contributes to understanding the behaviors of economic agents on digital platforms in situations where ratings and reviews alone may not sufficiently mitigate risks associated with information asymmetry. It studies these issues in two different online markets: the short-term rental market and the online labor market for crowd-sourced data work.

The first chapter provides empirical evidence of moral hazard at the end of sellers' career on digital platforms. Using data from Airbnb, it examines how end-game considerations affect sellers' effort decisions. Leveraging a regulation on short-term rentals in the City of Los Angeles, the study identifies hosts anticipating their imminent exit from the platform due to non-compliance with new eligibility rules. The focus is on hosts who left the platform as a result of the regulation, measuring their effort through listings' ratings in effort-related categories such as check-in, cleanliness, and communication. Employing a Difference-in-Differences and Event Study approach, the study compares how effort-related ratings of listings changed compared to ratings on location after the regulation announcement and during its implementation. The findings uncover a notable decline in effort-related ratings during the final periods of hosts' career, indicating the potential presence of moral hazard as sellers near their exit from the platform. This suggests that rating systems, particularly those reliant on simple averaging of scores, might not adequately address information asymmetry risks in the latter stages of sellers' career.

The second chapter explores the strategies adopted by firms outsourcing AI related tasks in a crowd-sourcing platform. Using data from a leading commercial online labor platform, the study examines demand volume and types of AI related jobs outsourced on the platform along with strategies employed by requesters in crowd-sourcing data training jobs. It highlights a growing demand for data work on the platform since 2019, covering a wide range of industries and sectors and suggests that privacy concerns could explain the limited request for annotation of sensitive data. A regression framework allows for the identification of specific factors that distinguish demand for data work from other

tasks. The higher targeting of demand towards predefined groups of contributors based on experience or geographic location, along with a larger rejection probability for data annotation tasks, underscores the importance of quality execution for firms outsourcing in this domain. This chapter's findings offer guidance for new requesters on crowd-sourcing platforms, outlining popular tools used to ensure the collection of quality output and inform the platform about the most common strategies adopted by their clients.

Finally, the third chapter introduces a methodology for quantifying the quality of data annotation tasks carried out on micro-tasking platforms. These platforms face quality challenges due to a Principal-Agent problem, stemming from low financial incentives. This issue arises because firms do not adequately monitor the quality of the work performed. Econometric reduced-form regressions are insufficient to study quality issues in such setting: indeed platform data indicates whether each task has been validated (and compensated) or rejected, the actual quality of executed tasks remains unobserved. To disentangle the factors influencing task rejection, specifically workers effort and firms investigation decisions, a structural approach is adopted. This involves modeling the simultaneous demand and supply of effort on the platform, considering the moderating influence of expectations from both sides. The model's parameters are estimated using proprietary data from a leading micro-tasking platform. Findings suggest that observed task rejection metrics underestimate the quality of task execution. Furthermore, the chapter discusses alternative incentives structures with counterfactual simulations. The simulation exercises reveal for instance that imposing a wage penalty for workers with rejected tasks could induce higher effort and require less monitoring by the outsourcing firm.

5 Résumé en Français

Cette thèse contribue à la compréhension des comportements des agents économiques sur les plateformes numériques dans des situations où les évaluations et les avis ("ratings and reviews") ne parviennent pas à atténuer suffisamment les risques associés à l'asymétrie de l'information. Les trois chapitres de la thèse étudient ces questions sur deux marchés différents: le marché de la location à court terme sur Airbnb et le marché du travail en ligne pour l'entraînement des données sur les plateformes de "crowd-sourcing".

Le premier chapitre présente des preuves empiriques de l'aléa moral chez les vendeurs en fin de carrière sur les plateformes numériques. À partir de données de la plateforme Airbnb, il examine comment l'anticipation d'un départ immédiat de la plateforme influence l'effort des hôtes. En se basant sur la réglementation des locations de courte durée à Los Angeles, l'étude identifie les hôtes anticipant leur départ imminent de la plateforme en raison du non-respect des nouvelles règles d'éligibilité. Nous mesurons l'effort des hôtes à travers les évaluations de leurs logements dans des catégories telles que l'enregistrement ("check-in"), la propreté du logement et la communication avec l'hôte. En utilisant une approche de Différence dans les Différences et d'Étude d'Événement, le chapitre compare l'évolution des évaluations liées à l'effort par rapport aux évaluations sur l'emplacement du logement, après l'annonce de la réglementation et pendant sa mise en œuvre. Les résultats révèlent une baisse statistiquement significative des évaluations liées à l'effort lors des dernières périodes de la carrière des hôtes, suggérant ainsi la présence potentielle d'un aléa moral lorsque les vendeurs se préparent à quitter la plateforme. Cela souligne que les systèmes d'évaluation, surtout ceux basés sur une simple moyenne des notes, pourraient ne pas suffisamment aborder ce problème dans les phases finales de la carrière des vendeurs.

Le deuxième chapitre explore le marché de l'entraînement des données sur les plateformes de "crowd-sourcing" (ou "micro-tâches"). En utilisant des données provenant d'une plateforme commerciale de travail en ligne, l'étude examine le volume de la demande ainsi que le contenu des tâches liées à l'intelligence artificielle externalisées sur la plate-

forme, ainsi que les stratégies employées par les employeurs pour garantir une exécution correcte des tâches. Les résultats mettent en évidence une demande croissante de "travail de données" sur la plateforme depuis 2019, couvrant un large éventail d'industries et de secteurs. Des analyses de régression permettent d'identifier les facteurs spécifiques qui distinguent la demande de travail de données des autres tâches sur la plateforme. L'orientation plus marquée de la demande vers des groupes prédéfinis de contributeurs, en fonction de leur expérience ou de leur situation géographique, ainsi qu'une probabilité de rejet plus élevée pour les tâches d'annotation de données, soulignent l'importance de la qualité d'exécution des tâches pour les entreprises externalisant dans ce domaine. Les résultats de ce chapitre offrent des conseils aux nouvelles entreprises qui souhaitent externaliser des tâches liées à l'intelligence artificielle sur les marchés de travail en ligne, en décrivant des méthodes utilisées par d'autres employeurs pour garantir la collecte de résultats de qualité. Ils indiquent également aux opérateurs des plateformes les outils privilégiés par leurs clients, qui peuvent être renforcés pour améliorer leur attractivité.

Enfin, le troisième chapitre présente une méthodologie visant à quantifier la qualité des tâches d'annotation de données effectuées sur des plateformes de micro-tâches. Ces plateformes sont confrontées à des problèmes de qualité en raison du problème "Principal-Agent", exacerbé par des faibles incitations financières. Ce problème découle du manque de contrôle adéquat de la qualité du travail effectué par les entreprises. Les méthodes économétriques traditionnelles en "forme-reduite" sont insuffisantes pour étudier les questions de qualité dans un tel contexte: alors que les données de la plateforme indiquent si chaque tâche a été validée (et rémunérée) ou rejetée, la qualité réelle des tâches exécutées reste inobservée. Pour séparer les facteurs influençant le rejet des tâches, notamment l'effort des travailleurs et les décisions d'investigation de l'entreprise, une approche structurelle est adoptée. Il s'agit de modéliser la demande et l'offre simultanées d'efforts sur la plateforme, en tenant compte de l'influence modératrice des attentes des deux parties. Les paramètres du modèle sont estimés avec de données exclusives provenant d'une plateforme de micro-tâches de premier plan. Les résultats suggèrent que les mesures de rejet des tâches observées sous-estiment la qualité de l'exécution des tâches. De plus,

le chapitre propose une correction basée sur le taux de surveillance de l'entreprise pour une estimation plus précise de la qualité des tâches, et discute des structures d'incitation alternatives avec des simulations contrefactuelles. Par exemple, les exercices de simulation révèlent que l'imposition d'une pénalité salariale aux travailleurs dont les tâches sont rejetées pourrait induire un effort plus important et nécessiter moins de vérification de qualité par l'employeur.

References

Acemoglu, Daron, and David Autor. 2011. "Skills, tasks and technologies: Implications for employment and earnings." In *Handbook of labor economics,* 4:1043–1171. Elsevier.

Agley, Jon, Yunyu Xiao, Rachael Nolan, and Lilian Golzarri-Arroyo. 2022. "Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7." *Behavior research methods* 54 (2): 885–897.

Agrawal, Ajay, Joshua S Gans, and Avi Goldfarb. 2019. "Artificial intelligence: The ambiguous labor market impact of automating prediction." *Journal of Economic Perspectives* 33 (2): 31–50.

Akerlof, George A. 1970. "The market for "lemons": Quality uncertainty and the market mechanism." *The Quarterly Journal of Economics,* 488–500.

Al Kuwatly, Hala, Maximilian Wich, and Georg Groh. 2020. "Identifying and measuring annotator bias based on annotators' demographic characteristics." In *Proceedings of the fourth workshop on online abuse and harms,* 184–190.

Albert, A., and J. A. Anderson. 1984. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." *Biometrika* 71 (1): 1–10.

Alekseeva, Liudmila, José Azar, Mireia Gine, Sampsa Samila, and Bledi Taska. 2021. "The demand for AI skills in the labor market." *Labour economics* 71:102002.

Anderson, Michael, and Jeremy Magruder. 2012. "Learning from the Crowd: Regression Discontinuity Estimates of the Effects of an Online Review Database." *The Economic Journal* 122 (563): 957–989.

Arhin, Kofi, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh. 2021. "Ground-Truth, Whose Truth?–Examining the Challenges with Annotating Toxic Text Datasets." *arXiv preprint arXiv:2112.03529.*

Arrow, Kenneth J. 1965. "Uncertainty and the Welfare Economics of Medical Care: Reply (The Implications of Transaction Costs and Adjustment Lags)." *The American Economic Review* 55 (1/2): 154–158.

———. 1986. "Chapter 23 Agency and the market," 3:1183–1195. Handbook of Mathematical Economics. Elsevier.

Bajari, Patrick, Han Hong, John Krainer, and Denis Nekipelov. 2010. "Estimating Static Models of Strategic Interactions." *Journal of Business Economic Statistics* 28 (4): 469–482.

Beck, Jacob. 2023. "Quality aspects of annotated data: A research synthesis." *AStA Wirtschafts-und Sozialstatistisches Archiv,* 1–23.

Beck, Jacob, Stephanie Eckman, Rob Chew, and Frauke Kreuter. 2022. "Improving Labeling Through Social Science Insights: Results and Research Agenda." In *International Conference on Human-Computer Interaction,* 245–261. Springer.

Bekkerman, Ron, Maxime C Cohen, Edward Kung, John Maiden, and Davide Proserpio. 2023. "The effect of short-term rentals on residential investment." *Marketing Science* 42 (4): 819–834.

Belleflamme, Paul, and Martin Peitz. 2018. "Inside the engine room of digital platforms: Reviews, ratings, and recommendations."

———. 2021. *The Economics of Platforms.* Cambridge University Press.

Belletti, Chiara, Daniel Erdsiek, Ulrich Laitenberger, and Paola Tubaro. 2021. *Crowdworking in France and Germany.* Technical report. ZEW Expert Brief.

Berg, Janine, Marianne Furrer, Ellie Harmon, Uma Rani, and M Six Silberman. 2018. "Digital labour platforms and the future of work." *Towards Decent Work in the Online World. Rapport de l'OIT.*

Berg, Janine, Uma Rani, et al. 2021. "Working conditions, geography and gender in global crowdwork." *Work and Labour Relations in Global Platform Capitalism. Cheltenham: Edward Elgar,* 93–110.

Brynjolfsson, Erik, and Tom Mitchell. 2017. "What can machine learning do? Workforce implications." *Science* 358 (6370): 1530–1534.

Cabral, Luis, and Ali Hortacsu. 2010. "The dynamics of seller reputation: Evidence from eBay." *The Journal of Industrial Economics* 58 (1): 54–78.

Carnehl, Christoph, Maximilian Schaefer, André Stenzel, and Kevin Ducbao Tran. 2022. "Value for money and selection: How pricing affects airbnb ratings." *Innocenzo Gasparini Institute for Economic Research Working Paper Series.*

Chandler, Dana, and Adam Kapelner. 2013. "Breaking monotony with meaning: Motivation in crowdsourcing markets." *Journal of Economic Behavior & Organization* 90:123–133.

Chevalier, Judith A, and Dina Mayzlin. 2006. "The effect of word of mouth on sales: Online book reviews." *Journal of marketing research* 43 (3): 345–354.

Corporaal, Greetje F, and Vili Lehdonvirta. 2017. "Platform sourcing: How Fortune 500 firms are adopting online freelancing platforms." *University of Oxford.*

Corrigan-Gibbs, Henry, Nakull Gupta, Curtis Northcutt, Edward Cutrell, and William Thies. 2015. "Deterring cheating in online environments." *ACM Transactions on Computer-Human Interaction (TOCHI)* 22 (6): 1–23.

Crémer, Jacques, Patrick Rey, and Jean Tirole. 2000. "Connectivity in the commercial Internet." *The Journal of Industrial Economics* 48 (4): 433–472.

Datta, Namita, Chen Rong, Sunamika Singh, Clara Stinshoff, Nadina Iacob, Natnael Simachew Nigatu, Mpumelelo Nxumalo, and Luka Klimaviciute. 2023. *Working Without Borders: The Promise and Peril of Online Gig Work.* Technical report. Washington, DC: World Bank.

Davani, Aida Mostafazadeh, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. "Hate Speech Classifiers Learn Normative Social Stereotypes." *Transactions of the Association for Computational Linguistics* 11:300–319.

Dellarocas, Chrysanthos. 2006. "Reputation mechanisms." *Handbook on economics and information systems,* 629–660.

Dellarocas, Chrysanthos, and Charles A. Wood. 2008. "The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias." *Management Science* 54 (3): 460–476.

Dube, Arindrajit, Jeff Jacobs, Suresh Naidu, and Siddharth Suri. 2020. "Monopsony in online labor markets." *American Economic Review: Insights* 2 (1): 33–46.

Duch-Brown, Nestor, Gomez-Herrera Estrella, Frank Mueller-Langer, and Songul Tolan. 2022. "Market Power and Artificial Intelligence Work on Online Labour Markets." *Research Policy.*

Duch-Brown, Néstor, Estrella Gomez-Herrera, Frank Mueller-Langer, and Songül Tolan. 2022. "Market power and artificial intelligence work on online labour markets." *Research Policy* 51 (3): 104446.

Eickhoff, Carsten. 2018. "Cognitive biases in crowdsourcing." In *Proceedings of the eleventh ACM international conference on web search and data mining,* 162–170.

Evans, David S. 2020. "Deterring bad behavior on digital platforms." *Available at SSRN 3455384.*

Excell, Elizabeth, and Noura Al Moubayed. 2021. "Towards equal gender representation in the annotations of toxic language detection." *arXiv preprint arXiv:2106.02183.*

Fan, Ying, Jiandong Ju, and Mo Xiao. 2016. "Reputation premium and reputation management: Evidence from the largest e-commerce platform in China." *International Journal of Industrial Organization* 46:63–76.

Farronato, Chiara, and Georgios Zervas. 2022. *Consumer reviews and regulation: evidence from NYC restaurants.* Technical report. National Bureau of Economic Research.

Figueroa, Rosa L, Qing Zeng-Treitler, Sasikiran Kandula, and Long H Ngo. 2012. "Predicting sample size required for classification performance." *BMC medical informatics and decision making* 12:1–10.

Fossen, Frank M, and Alina Sorgner. 2019. "New digital technologies and heterogeneous employment and wage dynamics in the United States: Evidence from individual-level data." *IZA Discussion paper.*

Fradkin, Andrey, Elena Grewal, and David Holtz. 2018. "The determinants of online review informativeness: Evidence from field experiments on Airbnb." *SSRN Electronic Journal* 41:1–12.

———. 2021. "Reciprocity and Unveiling in Two-Sided Reputation Systems: Evidence from an Experiment on Airbnb." *Marketing Science* 40 (6): 1013–1029.

Frank, Morgan R, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, et al. 2019. "Toward understanding the impact of artificial intelligence on labor." *Proceedings of the National Academy of Sciences* 116 (14): 6531–6539.

Gibbons, Robert, and Kevin J Murphy. 1992. "Optimal incentive contracts in the presence of career concerns: Theory and evidence." *Journal of Political Economy* 100 (3): 468–505.

Hirth, Matthias, Thobias Hoßfeld, and Phuoc Tran-Gia. 2011. "Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com." *in Proceedings of the Workshop on Future Internet and Next Generation Networks.*

Hirth, Matthias, Tobias Hoßfeld, and Phuoc Tran-Gia. 2013. "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms." *Mathematical and Computer Modelling* 57 (11-12): 2918–2932.

Ho, Chien-Ju, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. "Incentivizing high quality crowdwork." In *Proceedings of the 24th International Conference on World Wide Web,* 419–429.

Holmström, Bengt. 1979. "Moral hazard and observability." *The Bell journal of economics,* 74–91.

———. 1999. "Managerial incentive problems: A dynamic perspective." *The review of Economic studies* 66 (1): 169–182.

Hornuf, Lars, Sonja Mangold, and Yayun Yang. 2023. "Players in the Crowdsourcing Industry." In *Data Privacy and Crowdsourcing: A Comparison of Selected Problems in China, Germany and the United States,* 5–18. Springer.

Horton, John J. 2010. "Online labor markets." In *International workshop on internet and network economics,* 515–522. Springer.

Huang, Fan, Haewoon Kwak, and Jisun An. 2023. "Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech." *arXiv preprint arXiv:2302.07736.*

ILO. 2022. *Global Wage Report 2020-21: Wages and minimum wages in the time of COVID-19,* December.

Ipeirotis, Panagiotis G, Foster Provost, and Jing Wang. 2010. "Quality management on amazon mechanical turk." In *Proceedings of the ACM SIGKDD workshop on human computation,* 64–67.

Irani, Lilly. 2015. "The cultural work of microwork." *New Media & Society* 17 (5): 720–739.

Jullien, Bruno, and Wilfried Sand-Zantman. 2021. "The economics of platforms: A theory guide for competition policy." *Information Economics and Policy* 54:100880.

Kässi, Otto, and Vili Lehdonvirta. 2018. "Online labour index: Measuring the online gig economy for policy and research." *Technological forecasting and social change* 137:241–248.

Kingsley, Sara, Mary-Louise Gray, and Siddharth Suri. 2014. "Monopsony and the crowd: labor for lemons?" *Available at SSRN 3257857.*

Klein, Tobias J, Christian Lambertz, and Konrad O Stahl. 2016. "Market transparency, adverse selection, and moral hazard." *Journal of political economy* 124 (6): 1677–1713.

Koster, Hans RA, Jos Van Ommeren, and Nicolas Volkhausen. 2021. "Short-term rentals and the housing market: Quasi-experimental evidence from Airbnb in Los Angeles." *Journal of Urban Economics* 124:103356.

Kretschmer, Martin, Tobias Kretschmer, Alexander Peukert, and Christian Peukert. 2023. "The risks of risk-based AI regulation: taking liability seriously." *arXiv preprint arXiv:2311.14684.*

Laitenberger, Ulrich, Steffen Viete, Olga Slivko, Michael Kummer, Kathrin Borchert, and Matthias Hirth. 2022. "Unemployment and Online Labor - Evidence from Microtasking." *MIS Quarterly.*

Larimore, Savannah, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. "Reconsidering annotator disagreement about racist language: Noise or signal?" In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media,* 81–90.

Le Ludec, Clément, Maxime Cornet, and Antonio A Casilli. 2023. "The problem with annotation. Human labour and outsourcing between France and Madagascar." *Big Data & Society* 10 (2): 20539517231188723.

Lee, Neil, and Stephen Clarke. 2019. "Do low-skilled workers gain from high-tech employment growth? High-technology multipliers, employment and wages in Britain." *Research Policy* 48 (9): 103803.

Li, Hui, Yijin Kim, and Kannan Srinivasan. 2022. "Market shifts in the sharing economy: The impact of airbnb on housing rentals." *Management Science.*

Liu, Yi, Pinar Yildirim, and Z John Zhang. 2021. "Social media, content moderation, and technology." *arXiv preprint arXiv:2101.04618.*

Luca, Michael. 2011. "Reviews, Reputation, and Revenue: The Case of Yelp.Com." *Harvard Business School NOM Unit Working Paper No. 12-016.*

Mason, Winter, and Duncan J. Watts. 2010. "Financial incentives and the 'performance of crowds'." *ACM SigKDD Explorations Newsletter.*

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier. 2014. "Promotional reviews: An empirical investigation of online review manipulation." *American Economic Review* 104 (8): 2421–2455.

Miceli, Milagros, and Julian Posada. 2021. "Wisdom for the crowd: discoursive power in annotation instructions for computer vision." *arXiv preprint arXiv:2105.10990.*

———. 2022. "The Data-Production Dispositif." *Proceedings of the ACM on Human-Computer Interaction* 6 (CSCW2): 1–37.

Miklós-Thal, Jeanine, and Hannes Ullrich. 2016. "Career prospects and effort incentives: Evidence from professional soccer." *Management Science* 62 (6): 1645–1667.

Muchnik, Lev, Sinan Aral, and Sean J Taylor. 2013. "Social influence bias: A randomized experiment." *Science* 341 (6146): 647–651.

OECD. 2021. *Measuring the Digital Transition: A Roadmap for the Future,* November.

Pauly, Mark V. 1968. "The economics of moral hazard: comment." *The american economic review* 58 (3): 531–537.

Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti. 2014. "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk." *Behavior research methods* 46:1023–1031.

Petrin, Amil, and Kenneth Train. 2010. "A Control Function Approach to Endogeneity in Consumer Choice Models." *Journal of Marketing Research* 47 (1): 3–13.

Poirier, Dale J. 1980. "Partial observability in bivariate probit models." *Journal of Econometrics* 12 (2): 209–217.

Richter, Aaron N, and Taghi M Khoshgoftaar. 2020. "Sample size determination for biomedical big data with limited labels." *Network Modeling Analysis in Health Informatics and Bioinformatics* 9:1–13.

Rivera, Emilio D, Benjamin M Wilkowski, Aaron J Moss, Cheskie Rosenzweig, and Leib Litman. 2022. "Assessing the Efficacy of a Participant-Vetting Procedure to Improve Data-Quality on Amazon's Mechanical Turk." *Methodology* 18 (2): 126–143.

Rochet, Jean-Charles, and Jean Tirole. 2003. "Platform competition in two-sided markets." *Journal of the european economic association* 1 (4): 990–1029.

Rogstadius, Jakob, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. "An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets." In *Proceedings of the international AAAI conference on web and social media,* 5:321–328. 1.

Rossi, Michelangelo. 2018. "Asymmetric information and review systems: the challenge of digital platforms." *Economic analysis of the digital revolution,* 47.

Rossi, Michelangelo. 2023. "Competition and Reputation in an Online Marketplace: Evidence from Airbnb." *Management Science.*

Shaw, Aaron D, John J Horton, and Daniel L Chen. 2011. "Designing incentives for inexpert human raters." In *Proceedings of the ACM 2011 conference on Computer supported cooperative work,* 275–284.

Smith, Scott M, Catherine A Roster, Linda L Golden, and Gerald S Albaum. 2016. "A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples." *Journal of Business Research* 69 (8): 3139–3148.

Stephany, Fabian, Michael Dunn, Steven Sawyer, and Vili Lehdonvirta. 2020. "Distancing bonus or downscaling loss? The changing livelihood of US online workers in times of COVID-19." *Tijdschrift voor economische en sociale geografie* 111 (3): 561–573.

Stuart, Mark, Simon Joyce, Calum Carson, Vera Trappmann, Charles Umney, Chris Forde, Liz Oliver, Danat Valizade, Kate Hardy, Gabriella Alberti, et al. 2017. *The social protection of workers in the platform economy.* Technical report. Publications Office of the European Union.

Thuan, Nguyen Hoang, Pedro Antunes, and David Johnstone. 2013. "Factors influencing the decision to crowdsource." In *Collaboration and Technology: 19th International Conference, CRIWG 2013, Wellington, New Zealand, October 30–November 1, 2013, Proceedings 19,* 110–125. Springer.

Train, Kenneth E. 2009. *Discrete Choice Methods with Simulation.* 2nd ed. Cambridge University Press.

Tubaro, Paola, and Antonio A Casilli. 2019. "Micro-work, artificial intelligence and the automotive industry." *Journal of Industrial and Business Economics* 46:333–345.

Tubaro, Paola, Antonio A Casilli, and Marion Coville. 2020. "The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence." *Big Data & Society* 7 (1): 2053951720919776.

Woodcock, Jamie, and Mark Graham. 2019. "The gig economy." *A critical introduction. Cambridge: Polity.*

Xu, Lei, Tingting Nian, and Luis Cabral. 2020. "What makes geeks tick? A study of stack overflow careers." *Management Science* 66 (2): 587–604.

Yoganarasimhan, Hema. 2013. "The value of reputation in an online freelance marketplace." *Marketing Science* 32 (6): 860–891.

Zervas, Georgios, Davide Proserpio, and John W Byers. 2021. "A first look at online reputation on Airbnb, where every stay is above average." *Marketing Letters* 32:1–16.

ECOLE DOCTORALE

Titre : Essais empiriques sur les asymétries d'information sur les plateformes numériques

Mots clés : Plateformes Numériques ; Asymétries d'Information ; Systèmes de réputation ; Travail en Ligne

Résumé : Cette thèse étudie les problèmes liés à l'asymétrie d'information dans les marchés numériques. Elle vise à comprendre le comportement des agents économiques lorsque les outils de réputation standard, tels que les évaluations des consommateurs, peuvent être insuffisants. Le premier chapitre examine l'impact des préoccupations de "fin de jeu" sur l'efficacité des systèmes de réputation dans les marchés numériques. En utilisant des données d'Airbnb, le chapitre analyse les décisions d'effort des hôtes anticipant leur départ en raison de la non-conformité à une réglementation de location à court terme dans la ville de Los Angeles. Avec une approche de Différence-en-Différences et une Étude d'Événement, nous comparons comment les notes liées à l'effort d'un hôte ont changé, par rapport aux notes sur l'emplacement de son logement, après l'annonce de la réglementation et pendant sa mise en œuvre. Les résultats révèlent une diminution statistiquement significative des notes liées à l'effort, soulignant les limites des systèmes de réputation pour aborder l'aléa moral dans un jeu fini. Les deuxième et troisième chapitres de cette thèse étudient le comportement des entreprises et des travailleurs sur une plateforme commerciale de travail à la tâche caractérisée par l'anonymat et des interactions limitées entre employeurs et employés. Le deuxième chapitre fournit des aperçus descriptifs sur la manière dont les plateformes de travail à la tâche sont utilisées pour externaliser des tâches liées à l'IA. L'étude commence par fournir du contexte sur la plateforme étudiée et dévoile une demande croissante pour des micro-tâches liées au travail de données. Le chapitre examine également comment les entreprises assurent la qualité d'exécution des tâches grâce à la sélection des travailleurs, à la fixation des salaires et à la surveillance. Un cadre de régression permet d'identifier les facteurs spécifiques qui distinguent la demande de travail de données des autres tâches. Le plus grand ciblage de la demande vers des groupes prédéfinis de contributeurs basés sur l'expérience ou la localisation géographique, ainsi qu'une probabilité de rejet plus élevée pour les tâches d'annotation de données, soulignent l'importance de la qualité d'exécution pour les entreprises externalisant dans ce domaine. Le dernier chapitre explore un problème Principal-Agent, affectant la qualité d'exécution des tâches d'annotation de données sur les plateformes de micro-tâches. Ce problème ne peut pas être adéquatement traité avec une approche de forme réduite. Il découle de la surveillance peu fréquente de la qualité du travail par les entreprises, favorisant l'aléa moral chez les travailleurs. Un modèle structurel évalue l'équilibre de l'offre et de la demande d'effort, révélant que les métriques reposant sur l'observation du rejet des tâches sous-estiment la qualité. L'étude propose une correction plus précise basée sur le taux de surveillance propre à l'entreprise et simule des régimes d'incitations contrefactuels. Les exercices de simulation révèlent qu'une pénalité salariale pour les travailleurs ayant des tâches rejetées pourrait induire un effort plus élevé et nécessiter moins de surveillance. Une approche alternative, bien que plus coûteuse pour la plateforme, pour encourager l'effort, implique de fournir des subventions pour les efforts de surveillance des entreprises.

Title : Empirical essays on information asymmetries on digital platforms

Keywords : Digital Platforms ; Information Asymmetries ; Reputation Systems ; Online Labor

Abstract : This thesis studies issues related to information asymmetry in digital markets. It aims at understanding the behavior of economic agents when standard reputation tools, such as ratings and reviews, may fall short. The first chapter investigates the impact of "end-of-game concerns" on the effectiveness of reputation systems as monitoring tools in digital markets. Using data from Airbnb, the chapter examines the effort decisions of hosts anticipating their exit due to non-compliance with a short-term rental regulation in the City of Los Angeles. With a Difference-in-Differences and Event Study approach, we compare how listing's effort-related ratings changed, compared to ratings on location, after the regulation announcement and during its implementation. The findings reveal a statistically significant decrease in effort-related ratings during the hosts' final periods, highlighting the limitations of reputation systems in addressing moral hazard within a finite game. The second and third chapters of this thesis study firm and worker behavior on a commercial crowd-working platform characterized by anonymity and limited employer-employee interactions. The second chapter provides descriptive insights into how crowd-working platforms are used for outsourcing AI-related tasks, particularly focusing on data training. The study begins by providing context on the platform under study and unveiling a recent growing demand for crowd-sourced data work. The chapter also examines how firms ensure tasks' execution quality through worker selection, wage setting, and monitoring. A regression framework allows for the identification of specific factors that distinguish demand for data work from other tasks. The higher targeting of demand towards predefined groups of contributors based on experience or geographic location, along with a larger rejection probability for data annotation tasks, underscores the importance of quality execution for firms outsourcing in this domain. The final chapter explores a Principal-Agent problem arising from monetary incentives, affecting the quality of the execution of data annotation tasks on crowd-sourced platforms. This problem cannot be adequately addressed with a reduced form approach. It stems from firms infrequently monitoring the quality of work, fostering moral hazard by the workers. A structural model assesses the equilibrium demand and supply of effort, revealing that metrics relying on observed task rejection underestimate quality. The study suggests a more accurate back-of-the-envelope correction based on a firm's own monitoring rate and simulates counterfactual incentive schemes. The simulation exercises reveal that a wage penalty for workers with rejected tasks could induce higher effort and require less monitoring. An alternative approach, although more costly for the platform, to encourage effort and reduce the likelihood of overlooking poor quality, implies providing subsidies for firms' monitoring efforts.

Institut Polytechnique de Paris
91120 Palaiseau, France