POMS PRODUCTION AND OPERATIONS MANAGEMENT

PRODUCTION AND OPERATIONS MANAGEMENT POMS

Vol. 0, No. 0, xxxx–xxxx 2021, pp. 1–18
ISSN 1059-1478 | EISSN 1937-5956 | 21 | 00 | 0001

DOI 10.1111/poms.13341


# When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform

Gordon Burtch*
Carlson School of Management, University of Minnesota, 321 19th Ave. S., Minneapolis, Minnesota 55455, USA, gburtch@umn.edu

Yili Hong
C.T. Bauer College of Business, University of Houston, 4750 Calhoun Rd, Houston, Texas 77204, USA, yilihong@uh.edu

Senthil Kumar
Humphrey School of Public Affairs, 301 19th Ave. S., Minneapolis, Minnesota 55455, USA, natar062@umn.edu

We consider the role of online dispute resolution (ex-post guarantees of supplier quality) when introduced in the presence of an online reputation system (an ex-ante informational mechanism), in the context of online service procurement platforms. We argue that dispute resolution will reduce buyers' reliance on reputation systems in their hiring decisions to varying extents, depending on the nature of the work required. We assess these predictions using proprietary data capturing projects, service providers, bids, and hiring decisions around a natural experiment: the introduction of a new dispute resolution system at a major online service procurement platform. We provide evidence consistent with our expectations; introducing a dispute resolution system led buyers to reduce their consideration of service provider rating volumes in hiring decisions, particularly for projects where service provider performance could be evaluated objectively by a third party (e.g., data entry, as opposed to more subjective, creative work, like logo design). We also report a variety of additional analyses, which demonstrate the robustness of our findings to alternative measures, dynamics of the effects depending on buyers' experience with the dispute process, and the impact of the dispute service on buyers' propensity to enter ratings of service providers. These findings provide empirical evidence that dispute resolution can be an effective, alternative means of mitigating supplier quality risks in online service procurement markets in place of ex-ante signals of provider quality. However, this is particularly true in settings where the output of work contracted can be objectively evaluated by a third party.

*Key words:* dispute resolution; reputation system; online service procurement; digital platform
*History*: Received: September 2019; Accepted: November 2020 by Subodha Kumar after 2 revisions

## 1. Introduction

Large-scale web-based service marketplaces facilitate semi-anonymous transactions between geographically dispersed buyers and service providers (Allon et al. 2017, Lin et al. 2018, Tunca et al. 2014). These platforms enable on-demand service provisioning among strangers, and they present unique issues that scholars of operations management and information systems have recently begun to consider (Chen et al. 2020, Hu 2020, Sun and Xu 2018, Taylor 2018; Benjaafar and Hu 2020). The semi-anonymous nature of such platforms, paired with their relatively low barriers to entry, raises the prospect that they may become a breeding ground for opportunistic behavior absent the design of appropriate mechanisms (Hu 2020).

In traditional offline markets, the primary approach to managing supplier risk has been through contract design (Lim 2001, Tang 2006). However, innovative, peer-to-peer online marketplaces bring with them novel issues. Most notably, these markets are characterized by scale, with platforms hosting myriad buyers and sellers at any given time. Furthermore, the market is characterized by the geographic separation of buyers and service providers. This latter aspect, paired with the comparatively smaller value of contracts, makes contract enforcement through traditional legal channels a difficult prospect. Accordingly, platforms must rely on alternative, home-grown mechanisms to facilitate efficient matching and mitigate supplier risks and market frictions. Platform operators commonly rely on informational mechanisms, such as online rating systems (Lin et al. 2018, Snir and Hitt 2003, Sun and Xu 2018), which employers can use to screen service providers. These rating systems typically enable a persistent reputation,

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

allowing market participants to provide peer feedback following each transaction (Dellarocas 2003).

Unfortunately, these informational mechanisms are not a panacea. First, online rating systems create a cold-start problem for newly registered service providers, who are unable to transfer their offline work experience and reputation to the online platform (Pallais 2014). Second, online rating systems are subject to various rating biases (Hu et al. 2017) and are known to suffer from systematic inflation (Filippas et al. 2019), two issues that reduce the informativeness of ratings. Finally, in the specific context of online service procurement platforms, a variety of transaction types are typically facilitated, so ratings of past performance often do not translate well across projects. Therefore, historical ratings often fail to provide an accurate signal of a service provider's future performance or fitness for a particular job (Kokkodis and Ipeirotis 2015).

A great deal of academic work has centered on addressing these issues, attempting to devise policies and features for rating systems that mitigate or account for each concern (e.g., Garg and Johari 2018, Kokkodis 2019, Kokkodis and Ipeirotis 2015, Nosko and Tadelis 2015). Relatedly, online service procurement platform operators often seek to bolster their rating systems, supplementing them with other processes and mechanisms that reinforce desirable behaviors through non-reputational means. For example, platform operators also often implement service provider skill screening exams (Allon et al. 2017). And, increasingly, they also offer their own ex-post enforcement mechanisms, in the form of dispute resolution systems (Brett et al. 2007), the focus of the present study.

While online reputation systems offer a combination of ex-ante information (useful for screening transaction partners) and an ex-post reputational sanctioning (useful for mitigating opportunism), both mechanisms depend on ratings being representative and valued by platform participants. Moreover, these mechanisms offer little recourse if one party decides they are not concerned about potential reputational damage, and thus begins behaving badly. A platform-hosted dispute resolution system thus appears to be a natural complement to a reputation system. A dispute system can help to resolve disagreements between buyers and services providers after a project begins (Hong et al. 2016). A small body of prior work has studied the role of dispute resolution systems in online markets. That work has explored the factors that correlate with individuals' likelihood of engaging in the dispute process following a transaction (Macinnes et al. 2005), and how patterns of behavior within the dispute process associate with dispute outcomes (Brett et al. 2007, Friedman et al. 2004). However, little prior work has examined how dispute resolution systems may interact with reputation systems (Bakos and Dellarocas 2011) to influence hiring behavior, and what work does exist has largely been theoretical in nature.

Furthermore, given the heterogeneous nature of projects that online service procurement platforms accommodate, the effectiveness of dispute resolution systems is somewhat unclear. For example, some projects are characterized by routine tasks (the outcomes of which bear objective quality), for example, data entry, whereas other projects comprise mainly of creative tasks (the outcomes of which bear subjective quality), for example, logo design. Importantly, a buyer's willingness to rely on ex-post dispute resolution is likely to vary with the objectivity (subjectivity) of project output quality. This is because the platform's arbitration decisions will be more predictable when all parties (including the arbitrator) can readily agree about the project requirements and the quality of project output.

Our goal here is thus to empirically validate some of the theoretical predictions of past work (i.e., Bakos and Dellarocas 2011), exploring transaction contexts wherein dispute resolution systems may be expected to yield greater or lesser effects. Specifically, we focus on how the introduction of a dispute resolution system in the presence of an existing reputation system will shift buyers' hiring calculus, particularly concerning their reliance on ex-ante signals of service provider quality, for example, reputation, under alternative project types. We address the following formal questions: *Does the introduction of a platform provided dispute resolution system attenuate buyers' attention to peer reputation in selecting a service provider? To what extent does the reduced importance of seller reputation in hiring decisions (in the presence of a dispute resolution system) depend on project characteristics, namely the degree to which the quality of work output will be subjective, e.g., as in creative tasks?*

We leverage proprietary data from a large online service procurement platform. We consider the platform's initial introduction of a dispute resolution system in November 2008, quantifying shifts in the association between service providers' reputation and their probability of being hired for a project, conditional on bidding. We contrast these relationships before and after the introduction of the new dispute resolution system, considering heterogeneity by project type. With regard to the latter, we use both category-based measures and machine learning-based measures to identify and compare service contracts characterized by objective vs. subjective output quality.

We report empirical evidence consistent with our expectations. We show that the introduction of a dispute resolution mechanism led buyers to reduce their

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

consideration of service provider ratings when making hiring decisions. More concretely, we observe that buyers' response to service provider ratings declined by more than 50%, and these effects manifested more strongly in work domains where an objective evaluation of project output by a third party was more plausible, for example, Data Entry, IT & Software. In contrast, we observe weaker, or null effects in work domains characterized by horizontal quality (fit and subjective quality), for example, Logo Design, Writing. We demonstrate that our results are robust to the use of alternative measures of project type, as well as a consideration of alternative ex-ante signals of worker quality (namely, skill certifications). Finally, additional analyses demonstrate that the introduction of the dispute service led to significant increases in buyer satisfaction.

Our findings bear important implications. Our work contributes to a recent operations management literature on the design of innovative online, peer-to-peer marketplaces (Benjaafar and Hu 2020, Chen et al. 2020, Hu 2020). Chen et al. (2020), Hu (2020), Benjaafar and Hu (2020), variously review the recent OM literature on this topic and propose research agendas for the field. Our work speaks directly to a number of the open questions and domains those scholars raise.

From a practical perspective, web-based service marketplaces constitute one of the fastest-growing forms of employment today. Over the period spanning 2016 and 2017, more than 57 million individuals freelanced in the United States. This phenomenon has created many micro-level operational problems that pertain to the hiring relationship and collaboration between the buyer and the service providers, as well as in cases of failed projects, the cost to arbitrate and litigate, calling for efficient and preferably automated solutions. Our finding that service provider reputation plays a lesser role in the presence of dispute resolution is, on the whole, desirable, in the sense that it implies mitigation of the cold-start problem (Pallais 2014), without any apparent cost to buyers, in terms of reduced satisfaction. Moreover, the cost born by the platform to operate this service is relatively small; Modria.com, a major provider of online dispute resolution systems to eCommerce sites, reportedly resolves 90% of all disputes filed without human involvement.[1] Chen et al. (2020) highlight a need to better understand how platform intermediaries can effectively reduce search frictions and uncertainty for buyers and service providers. Similarly, Hu (2020) points to open research questions related to the understanding of platform mechanisms for managing information asymmetry between buyers and sellers. Our work speaks directly to such calls.

## 2. Related Work and Conceptual Background

### 2.1. Managing Supplier Quality Risk

An extensive literature in operations management has dealt with supplier risk, and effective approaches to its management (Tang 2006). The vast majority of that literature has focused on optimal contract formulation. For example, Swinney and Netessine (2009) study how buyers' contract duration preferences vary in the presence of possible supplier default, as well as the extent to which these contracts can effectively coordinate the supply chain. Lim (2001) presents a more directly relevant example, considering contracting as a means of managing information asymmetry around supplier quality. He considers inspection schemes (with a refund provided to the buyer when the quality of supply turns out to be poor) and compares them with warranty schemes, discussing whether and when each will be effective. To some extent, the dispute resolution system we consider here is analogous to an inspection scheme. Our work furthers the notion that the feasibility of inspection schemes depends on the characteristics of the supplied services, in platforms wherein services provided are heterogeneous.

In recent years, this literature has expanded in scope to include service procurement via innovative online marketplaces, such as the one we study here. Online platform markets differ from traditional supply chains because "the independent sellers in some marketplaces have the full right to set their own prices" though "the platform can … influence the matching process between buyers and sellers, e.g., by manipulating information disclosed to an interested party." (Hu 2020). One illustrative example is the provision of service provider capacity information. Horton (2019) considers how a platform intermediary in a web-based service marketplace may intervene to manage congestion through the revelation of service provider (seller) capacity information, that is, indicating how "busy" a service provider is likely to be. Absent such information, buyers or sellers may waste time and effort attempting to match with unresponsive parties, thus disclosing this information reduces those wasted efforts.

Matching may also be inefficient if a buyer transacts with an unexpectedly low-quality service provider. Web-based service procurement platforms are generally characterized by their scale, the semi-anonymous nature of transactions, and geographic separation between buyers and service providers. For these reasons, contract enforcement through traditional legal channels is a difficult proposition. As such, the design and implementation of effective mechanisms for

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

managing information asymmetry problems, and for facilitating efficient matching, are crucial issues for platform operators. Our thus work speaks to an as-yet understudied means by which platform operators can mitigate information asymmetry and enable buyers to manage supplier risks. Whereas past work has focused almost exclusively on ex-ante information mechanisms, such as reputation systems (Sun and Xu 2018) and skill screening exams (Allon et al. 2017), we consider an ex-post mechanism in the form of a platform-provided dispute resolution system.

More broadly, our work contributes to this prior literature, highlighting the need for more research that considers the role of novel mechanisms for managing supplier quality risk in online marketplaces. In particular, our work highlights the potential oversight role that an online platform operator can provide by implementing dispute resolution systems. Furthermore, our work highlights that dispute resolution systems can effectively mitigate supplier quality information asymmetry under certain conditions.

### 2.2. Managing Supplier Quality Risk for Online Services

Managing supplier quality risk is inherently challenging when services and transactions take place entirely online. Trust is critical to the sustained operation of online platform markets because employers and service providers both operate semi-anonymously (Ba and Pavlou 2002, Einav et al. 2016, Fugger et al. 2019, Hong and Shao 2020). Trust within online marketplaces is facilitated through the use of several institutional trust mechanisms (Pavlou and Gefen 2004), among which informational mechanisms for the ex-ante screening of transaction partners, for example, online feedback and reputation systems, are perhaps most prevalent (Dellarocas 2003). The impact of online feedback and reputation systems on the selection of products, services, and transaction partners has been studied in a variety of contexts and consistently found to have both statistically and economically significant impacts. Research on web-based service platforms has found that service providers who hold a higher rating exploit it to their benefit, in two ways (Moreno and Terwiesch 2014). First, high-rated service providers may raise their bid prices and charge a premium when engaging in the same volume of work. Second, they may maintain their bid price and pursue more project work. Between the two behaviors, service providers appear to primarily engage in the latter.

Although reputation mechanisms facilitate transactions in a variety of online markets, they nonetheless suffer from some problems (Tadelis 2016) and their effectiveness is limited in some scenarios (Lin et al. 2018). For example, Wang et al. (2017) document that, even when ratings are informative, they are only effective at inducing good behavior when platform participants plan to build their reputation for the long term. If long-run reputation is not a concern, agents are likely to engage in opportunistic behavior. Furthermore, the efficacy of these systems depends a great deal on the nature of the services transacted. For instance, Sun and Xu (2018) consider markets for services where production is collaborative, and thus where the quality of output is a function of the effort invested by both service providers and buyers. In such settings, those authors demonstrate that rating systems are less useful because ratings fail to provide a reliable signal of service provider quality. Notably, some services offered through online service procurement marketplaces require more collaboration between the service provider and the buyer. Most notably, some types of work, such as software development, are a result of significant cooperation between buyers and service providers (Hong et al. 2016), and they depend a great deal on the quality of requirement elicitation and application specification (Hong and Pavlou 2017).

A stream of work has also documented evidence of reporting biases in online feedback (Hu et al. 2009). Consumers are systematically more likely to report feedback when they have extreme experiences, for example (Hu et al. 2017). Beyond this, depending on the design of reputation mechanisms, reciprocity and retaliation may be a concern as well. Ye et al. (2014) provided evidence of strategic rating behavior, based on a natural experiment at eBay. Those authors showed that sellers often retaliate when they receive a negative evaluation by giving a lower rating to the buyer. This behavior results in rating inflation, as individuals either provide positive assessments or no evaluation at all. Indeed, rating inflation occurs in a variety of popular online platforms, including Airbnb (Fradkin et al. 2018, Zervas et al. 2015) and Upwork (Filippas et al. 2019).

Online reputation mechanisms can also lead to a cold-start problem (Pallais 2014). Service providers who lack reputation typically have a difficult time attracting buyers, and thus they have a difficult time building reputation, yielding a chicken-and-egg problem. In a field experiment conducted by Pallais on oDesk (now Upwork), the author contracted with inexperienced service providers and then supplied feedback on the output of a randomly chosen subset. She found that, 2 months after the experiment, the earnings of these service providers tripled as a result of the rating and review received, implying that buyers overlook many high-quality service providers as a result of over-reliance on reputation mechanisms. Thus, online reputation systems can drive a great deal of inefficiency in these platforms.

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

A potentially useful approach to mitigating the cold-start problem is for platforms to offer certification exams to assess worker skills and ability. An example of work in the operations management literature considering the value of certifications in online service procurement platforms is that by Allon et al. (2017). Those authors consider the implementation of certification exams, studying the optimal number of evaluations required to maximize platform revenue. They conclude that as the fit between worker skills and customer needs lessens, or as customer processing needs become more variable, skill screening exams can deliver significant improvements in platform revenue. Worker certification exams offer platform operators a more direct line of control over platform activity, unlike online rating systems, which are primarily user driven. Importantly, however, skills certifications specifically facilitate the ex-ante screening of transaction partners; they do little to ensure desirable behavior upon contracting. In contrast, dispute resolution systems provide platform operators with a similar means of directly enabling trust on the platform, providing an ex-post means of fostering desirable project outcomes. That said, the degree to which this is true is likely to vary with the characteristics of the job posting, as we discuss next.

### 2.3. Theorizing the Effects of a Dispute Resolution System

As highlighted above, reputation typically plays a vital role in the online hiring process (Hong and Pavlou 2017, Moreno and Terwiesch 2014), facilitating the ex-ante screening of transaction partners and providing an ex-post means of fostering good behavior, due to reputational concerns. Because transactions on these platforms are semi-anonymous and parties reside in different legal jurisdictions (e.g., other countries or states), legal recourse is generally not a viable option. Accordingly, reputation mechanisms provide critical information that buyers can use to infer supplier quality, based on past performance (Banker and Hwang 2008).

Although reputation systems are by far the most commonly studied informational mechanism employed by online platforms, dispute resolution systems are also commonly employed in practice, and they can also be an effective mechanism for fostering trust and good behavior. When buyers and service providers experience disagreements after a project begins, a dispute resolution system allows either party to lodge a complaint with the platform against the other party, and to submit evidence in support of their case. The implementation of a platform-provided dispute resolution system thus creates a new avenue for ex-post issue resolution that buyers may leverage.

Prior studies on dispute resolution systems have been primarily exploratory in nature, focusing on the process of dispute and how parties engage with the system. For example, Macinnes et al. (2005) identified factors that correlate with an individual's likelihood of initiating dispute resolution at eBay, considering characteristics of the individual, as well as the transaction. Friedman et al. (2004) considered individual behavior within a dispute interaction, such as the display of anger, and the influence on dispute outcomes. Brett et al. (2007) studied the relationship between the likelihood of successful settlement in eBay disputes and involved parties' use of language that attacks (or gives "face" to) the transaction partner. To our knowledge, no prior work has empirically examined how dispute resolution systems may combine with reputation systems to affect buyers' behaviors.

The procedural fairness of a dispute resolution system is a direct determinant of impersonal trust (Culnan and Armstrong 1999), thereby easing buyers' concerns that a service provider may engage in fraud or slack off once hired. As such, we expect a substitution effect between the dispute resolution system and the reputation system, such that the implementation of a dispute reputation system will lead to a general reduction in the importance of service provider reputation in buyers' hiring decisions. Beyond the general effect, however, we also expect that the strength of the effect will vary depending on the degree to which employers trust the dispute process to yield a fair outcome.

As noted earlier, this predictability of the dispute outcome will depend to a large extent on project characteristics. As with physical products, the quality of service output from a provider can be characterized based on the degree to which it varies horizontally, that is, subjective quality, or fit, vs. vertically, that is, objective quality (Bommer et al. 1995). In some types of work, it is relatively easy to measure job performance using objective criteria (Gottfredson 1991). Routine work, in particular, offers well-codified job requirements (Autor et al. 2003), which enable objective evaluation criteria. Based on the classifications developed by Autor and Dorn (2013), some tasks in online service procurement platforms are routine, namely data entry and software development. With routine tasks, the effort levels invested by a service provider will map more deterministically to work output. With these routine jobs, provider performance can be quantified directly via objective quality measures, for example, error rate, time to completion. Accordingly, when a significant discrepancy occurs between the expected and delivered quality, a buyer can expect a favorable outcome from a dispute resolution system. The impact of introducing dispute resolution on buyers' attention to service provider reputation is thus likely to be quite strong for this type of work, given outcomes are readily measurable.

By contrast, in project categories that are creativity oriented, such as creative writing and logo design, subjective preferences play a more significant role. In such domains, buyers are much less likely to trust that a dispute resolution system will side in their favor when a service provider's deliverable falls short of expectation. The quality of creative work product is perhaps best described in terms of Hotelling preferences (1929); that is, quality perceptions for creative work output are heavily dependent on subjective tastes. Although the buyer may feel work product is of low quality, that may not be readily apparent to a platform arbitrator. Indeed, the work requirements are typically harder to codify for creative oriented tasks (Autor et al. 2003). As such, we predict that the introduction of dispute resolution will be much less likely to change buyers' hiring behavior in creative work domains.

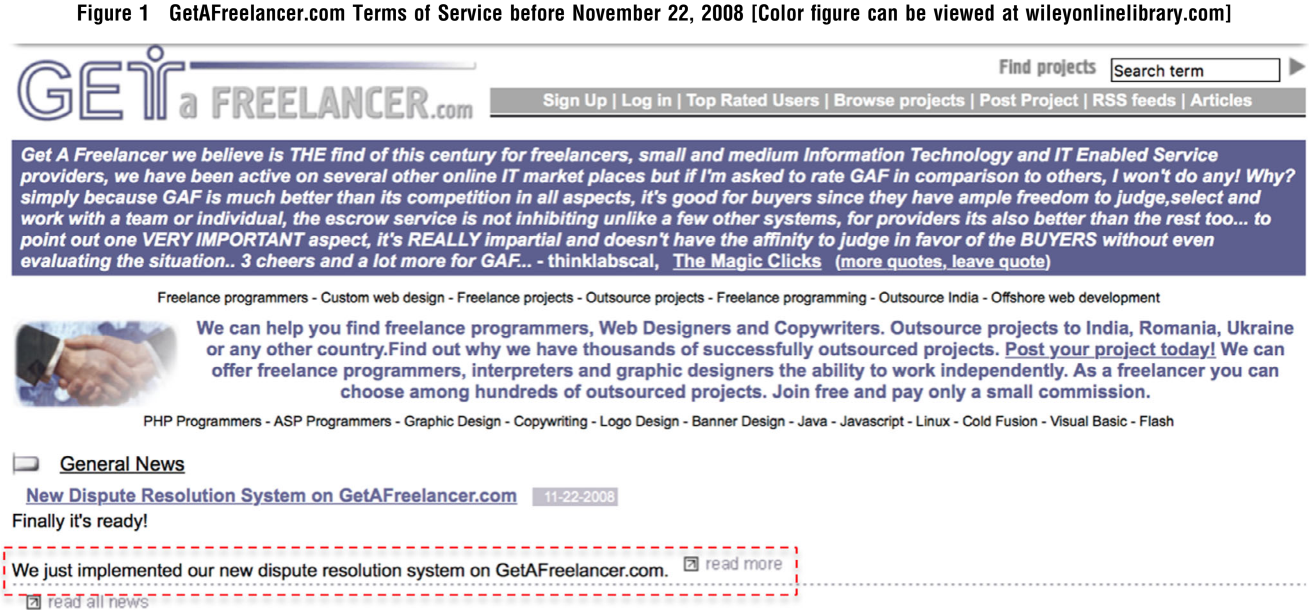
## 3. Study Context

We leverage proprietary data from a large online service procurement platform.[2] The natural experiment we consider is the introduction of a formal dispute resolution system on the platform, following its announcement on November 22, 2008. Before the introduction of the dispute resolution system, buyers and service providers on the platform had no official recourse for resolving disagreements with transaction partners. The platform's terms of service stated simply that the platform bore no liability or responsibility to the agreements and disputes between service providers and their buyers for contracted projects (see Figure 1). With the announcement of the new dispute resolution system, however, the platform updated its terms of service and advertised the change prominently on the website homepage. Figure 2 provides a screenshot of the website announcement on November 22, 2008. We also provide the details of the announcement related to the new dispute resolution system below for readers' reference.

Our estimation sample includes data on all projects in the four largest categories posted to the platform over the period between July of 2008 and May of 2009, which encompasses the date of the dispute service's introduction; that is, approximately 5 months before and 6 months following the change (November 22, 2008). Our primary analysis considers bids and hiring outcomes associated with projects posted under these categories on the platform: Writing, Design, IT & Software, and Data Entry. These categories collectively account for 80% of all project work on the platform. Thus, these categories capture the vast majority of market activity. Moreover, and most importantly, they provide a useful contrast in terms of the objectivity of associated work output quality.

For each project, we observe the information of the buyer, the project description, its budget, bids received, associated service providers' information, and the service provider selected for the contract. We also observe the characteristics of buyers. For example, we observe each service provider's past rating volume, whether he or she has any platform-provided certifications, his or her location, and various other aspects of his or her account profile. Leveraging this information, we consider potential changes in buyers' hiring behavior, which may be both beneficial and detrimental from the platform operator's perspective.

**Figure 1** GetAFreelancer.com Terms of Service before November 22, 2008 [Color figure can be viewed at wileyonlinelibrary.com]




Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

**Figure 2 Dispute Resolution System Announcement on November 22nd, 2008 [Color figure can be viewed at wileyonlinelibrary.com]**

General News

11-22-2008 **New Dispute Resolution System on GetAFreelancer.com**

Finally it's ready!

We just implemented our new dispute resolution system on GetAFreelancer.com. We beleive this will be a great tool that will help users find a solution to their disputes.

This is shortly how the new system works:

Round 1 - Automated Resolution

GetAFreelancer.com's automated resolution process allows the Parties to settle disputes through the GetAFreelancer.com's Web Site software. The software compares the Parties' offers and demands on a round-by-round basis. Each Party will have 3 opportunities to settle the dispute. If no Bids match after 3 rounds the software will suggest a solution ("Auto-offer") based on the Parties Bids.

Round 2 - Submit Evidences and Arbitration Fee Estimates

The Parties have 7 days to submit/upload all the material ("Evidences") the Parties want GetAFreelancer.com to consider if the dispute has to be resolved through Arbitration. GetAFreelancer.com will show the estimated Arbitration Fee (GetAFreelancer.com charge 2% or USD 10, whichever is greater, from Escrowed Funds for estimating the Arbitration Fee). Within 3 days each Party must either (i) accept one of the opponents previously submitted Bids, (ii) accept GetAFreelancer.com's Auto-offer, or (iii) agree to proceed to Arbitration.

Round 3 - Arbitration

GetAFreelancer.com shall serve as the arbitrator in a dispute between the Parties that have elected to use the Escrow Service and requested that a dispute shall be addressed through the Dispute Resolution Services. GetAFreelancer.com shall render its decision within 14 days after the submission of a dispute to arbitration. During this time, the Parties are encouraged to continue to negotiate an amicable settlement.

Please read our new Terms of Service carefully to fully understand all details on how the new dispute resolution system works:

http://www.getafreelancer.com/page.php?p=info/terms

Please note that as a user on GetAFreelancer.com you also automatically accept our new Terms of Service.

Have a Great Weekend!

Our analyses primarily focus on buyers' hiring behavior. Specifically, we consider how the introduction of the dispute resolution system in November of 2008 shifted the degree to which ratings and certifications factored into buyers' hiring decisions across projects of different types. That is, we consider the degree to which the presence of a dispute resolution system reduces buyers' concerns about or attention to information on service providers' quality, in general, as well as heterogeneity in that effect across types of projects. To evaluate this, we construct a panel of project-bid observations, incorporating reputation and certification information associated with each service provider placing a bid. Our outcome of interest in these estimations is a binary indicator of whether the service provider receives the project contract.

A key feature of our proprietary data set is that we observe a buyer's entire consideration set of service providers who submitted bids for the projects, free of measurement error. That is, our data generation process can reliably recover the actual scenario of each buyer's hiring decision. Moreover, we observe the same service providers' bids, across projects, over time, under differing reputation (as the service providers accrue more ratings). This set-up allows us to accurately estimate buyers' preferences over service provider characteristics across contexts, controlling for any attributes of a service provider or project (and buyer) that does not vary within the sample.

## 4. Data

We present definitions and descriptive statistics for the variables in our sample, in Tables 1 and 2. Writing projects involve requests for blog content, creative writing, travel writing, screenwriting, and so forth, while design work includes requests for animation, graphic design, website design, fashion design, illustration, etc. Broadly, these two categories represent the most artistic or creative projects in the market, where the quality of output is dependent upon subjective Hotelling (1929) preferences. In contrast, data entry projects involve readily verifiable output. Examples of such projects include requests to scrape and record data from websites, process orders, perform and record results of web searches, author Excel macros. IT & Software projects are software development tasks. Although these projects tend to be more

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

**Table 1 Variable Definitions**

| Variable | Definition |
|---|---|
| Hired | A binary indicator of whether the service provider was hired following his or her bid. |
| BidAmount | The bid amount entered by the service provider. |
| BidSequence | The service provider's position in the bidding sequence. |
| Ratings | The volume of ratings the service provider had received from prior buyers. |
| Objective | A binary indicator of whether the project was posted to the Data Entry or IT & Software project categories. |
| Language | A binary indicator of whether the service provider's country of residence and the buyer's country of residence share an official language. |
| Distance | The physical distance, in kilometers, between the service provider's self-reported location and that of the buyer. |
| HasCert | A binary indicator of whether the service provider holds any certifications. |
| EPP | A binary indicator of whether the buyer ultimately rated the service provider positively or not (9/10 or 10/10). This indicator equals 0 if no rating was provided, or a rating below nine was provided. |
| TurkObjective | A binary indicator predicted based on a random forest model trained with Amazon Mechanical Turk workers' hand-coded labels. This label is increasingly positive when a project's output (based on its textual description) was deemed more suitable for objective metric-based performance evaluation than subjective, opinion-based evaluation (e.g., a project manager's opinion). |
| HasMadeComplaint | A binary indicator of whether the buyer has ever previously entered a complaint against a service provider via the dispute resolution system. |
| HasReceivedComplaint | A binary indicator of whether the buyer has ever previously received a complaint from a service provider via the dispute resolution system. |

involved than data entry jobs, they are nonetheless well codified, typically being based on a required software specification, with well-defined deliverables and testing procedures.

## 5. Analysis and Results

### 5.1. Main Analyses

We model the hiring outcome as a function of service providers' rating volumes at the time of their bids (and subsequently as a function of whether a service provider holds a certification). We also include a service provider fixed effect, a project fixed effect (which subsumes the buyer fixed effect), and time dummies reflecting project posting times. We focus on reputation volume, and not valence because the reputation system suffers from extensive inflation, such that more than 95% of service providers in the platform have an average rating above 9 (out of 10), a phenomenon common to many online rating systems studied in the literature (Nosko and Tadelis 2015, Filippas et al. 2019). That said, buyers still reference a service provider's rating volumes when making a hiring decision, because these nonetheless speak to service provider experience and past buyers' willingness to hire the individual (Pallais 2014).

We begin by assessing general trends in buyers' reliance upon service provider ratings in their hiring decisions. We then progress to more nuanced regression specifications, which enable us to evaluate changes in employers' reliance on service provider ratings as a function of the work to be performed. We focus on the effect of ratings over time, before and after the introduction of the dispute resolution system. Moreover, we focus on how those dynamic effects differ as a function of project type, with some projects being highly objective in their requirements, for example, data entry, and others being highly subjective, for example, logo design. To conduct our analysis, we construct a binary indicator of whether work is *Objective* (vs. Subjective), and we interact it with our time dummies, treating the period immediately before the introduction of the dispute resolution system as our reference period. In this way, we can show that rating volumes play a stable role in the hiring decision over the period before the introduction of the dispute service, yet that they began to play a weaker role afterward.

Moreover, interacting the *Objective* indicator with our time dummies, we implement a difference-in-differences specification to demonstrate that the change in rating consideration by buyers over time, across project categories, is more considerable for *Objective* project work, as compared to *Subjective* project work. The equation below reflects our interacted regression specification, which we implement as a Linear Probability Model (LPM).[3] Specifically, our difference-in-differences estimation distinguishes changes in buyer response to service provider ratings over time, across project types.

$$\begin{aligned} Hired_{w,p} =& \beta_1 \cdot \ln(Ratings_{w,p}) + \beta_2 \cdot \ln(Ratings_{w,p}) \\ & \cdot Objective_p + \beta_3 \cdot \ln(Ratings_{w,p}) \cdot \sum_t \tau_p \\ & + \beta_4 \cdot \ln(Ratings_{w,p}) \cdot Objective_p \cdot \sum_t \tau_p \\ & + C_{w,p} + \alpha_w + \gamma_p + \varepsilon_{w,p} \end{aligned} \tag{1}$$

In equation (1), $w$ indexes service providers, $p$ indexes projects, $\alpha$ is a vector of service provider fixed effects, $\gamma$ is a vector of project fixed effects,[4] and $\tau$ is a vector of time (quarter) dummies. *Ratings* refers to the volume of prior ratings a service provider has obtained

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

**Table 2 Descriptive Statistics**

| Variable | Mean | Std. Deviation | Min | Max | Observations |
|---|---|---|---|---|---|
| Hired | 0.06 | 0.230 | 0.00 | 1.00 | 595,410 |
| BidAmount | 207.24 | 1,803.10 | 30.00 | 999,999.00 | 595,410 |
| BidSequence | 20.67 | 30.07 | 1.00 | 616.00 | 595,410 |
| Ratings | 32.16 | 100.87 | 0.00 | 972.00 | 595,410 |
| Objective | 0.41 | 0.49 | 0.00 | 1.00 | 595,410 |
| Language | 0.297 | 0.32 | 0.00 | 1.00 | 560,604 |
| Distance | 4,974.47 | 2,790.09 | 0.00 | 12,274.45 | 588,472 |
| HasCert | 0.26 | 0.44 | 0.00 | 1.00 | 595,410 |
| EPP | 0.55 | 0.50 | 0.00 | 1.00 | 33,344 |
| TurkObjective | 51.61 | 5.69 | 25.85 | 78.99 | 595,410 |
| HasMadeComplaint | 0.02 | 0.14 | 0.00 | 1.00 | 595,410 |
| HasReceivedComplaint | 0.01 | 0.10 | 0.00 | 1.00 | 595,410 |

at the time of their bid and *Objective* is our treatment group indicator, a binary measure of whether the project involves data entry, IT or software coding work (vs. design or writing work), and $C$ is a set of controls that the prior literature shows to be relevant (Hong and Pavlou 2017), including the physical distance between the service provider and buyer, shared language, the service provider's bid price, and the service provider's arrival sequence among project bidders. We employ robust standard errors in all models. The coefficients of interest here are $\beta_3$, which captures how the effect of ratings on hiring outcomes for *Subjective* work varies over time, and $\beta_4$, which reflects our difference-in-differences estimates, capturing how the influence of *Ratings* over time, differs for *Objective* projects, as compared to *Subjective* projects.

We first report the average effect of ratings on hiring, over time, in Table 3. Column 1 reports a model incorporating only a project (buyer) fixed effect, while column 2 reports results incorporating both project and service provider fixed effects. Subsequently, we report heterogeneous results across project types in Table 4. In Column 1 we include only a project fixed effect. In Column 2 we include both a project and service provider fixed effect. Because there is little variation in rating volume across time within providers, we would caution the reader to interpret the estimated main effect of rating volumes in the latter model. Broadly, we observe evidence consistent with our predictions. The effect of service providers' ratings on their probability of being hired was attenuated following the introduction of a dispute resolution system. Moreover, this effect was significantly larger for *Objective* projects.

Considering the specific coefficients, we first find that the effect of ratings is systematically more positive for Objective projects, in general, as compared to Subjective projects. In the period before the dispute resolution arrives, a 1% increase in the volume of ratings a service provider has previously received is associated with an approximate 0.6% *larger* increase in his or her probability of being hired for an Objective task, as compared to a *Subjective* task. Notably, this result is consistent with Kokkodis and Ipeirotis's (2015) finding regarding reputation transferability (feedback on past subjective work receives less consideration). Considering how this difference varies over time, with the introduction of the dispute resolution system, we see that the difference weakens by approximately half.

In terms of control variables, we observe that larger bid amounts (prices) are negatively associated with hiring, as expected, such that a 1% increase in the amount bid is associated with a 4.3% decline in the probability that a service provider is hired. Furthermore, we observe that service providers who bid later are more likely to be hired, perhaps because these service providers have the benefit of seeing competitors' pricing and optimize their bids on that information. We also observe that spatially distant service providers are systematically less likely to be hired. In contrast, service providers who share a common language with the buyer are more likely to be hired, all of which is consistent with prior work (Hong and Pavlou 2017). These results thus collectively provide support for our various predictions. In the next

**Table 3 Effect of Ratings Over Time (LPM /w FEs)**

| Variable | (1) | (2) |
|---|---|---|
| Ln(Ratings) • $t-2$ | 0.001 (0.001) | 0.001 (0.001) |
| Ln(Ratings) • $t-1$ | — | — |
| Ln(Ratings) • $t+0$ | −0.002*** (0.0006) | −0.002** (0.0007) |
| Ln(Ratings) • $t+1$ | −0.0017** (0.0006) | −0.0015* (0.0017) |
| Ln(Ratings) | 0.021*** (0.0004) | 0.001 (0.001) |
| Ln(BidAmount) | −0.039*** (0.001) | −0.043*** (0.001) |
| Ln(BidSequence) | 0.003*** (0.0003) | 0.002*** (0.001) |
| Language | 0.027*** (0.0015) | 0.022*** (0.003) |
| Ln(Distance) | −0.002*** (0.0002) | −0.002*** (0.0003) |
| Observations | 560,529 | 526,754 |
| Project FEs | Yes | Yes |
| Service Provider FEs | No | Yes |
| Adj. $R^2$ | 0.048 | 0.068 |

*Notes*: Robust standard errors in parentheses; Observations reduced in Model 2 due to loss of singleton service providers.
$^{***}p < 0.001$ ,$^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

**Table 4 Effect of Ratings on Hiring Over Time by Project Objectivity (LPM /w FEs)**

| Variable | (1) | (2) |
|---|---|---|
| Ln(Ratings) • $t-2$ | 0.0002 (0.001) | 0.001 (0.001) |
| Ln(Ratings) • $t-1$ | — | — |
| Ln(Ratings) • $t+0$ | −0.0014* (0.0007) | −0.001 (0.001) |
| Ln(Ratings) • $t+1$ | −0.001 (0.001) | −0.0002 (0.001) |
| Ln(Ratings) • Objective • $t-2$ | 0.0016 (0.0013) | 0.0017 (0.0015) |
| Ln(Ratings) • Objective • $t-1$ | — | — |
| Ln(Ratings) • Objective • $t+0$ | −0.003* (0.001) | −0.003* (0.001) |
| Ln(Ratings) • Objective • $t+1$ | −0.004** (0.001) | −0.004** (0.001) |
| Ln(Ratings) | 0.018*** (0.0005) | −0.003* (0.001) |
| Ln(Ratings) • Objective | 0.010*** (0.001) | 0.006*** (0.001) |
| Ln(BidAmount) | −0.039*** (0.001) | −0.043*** (0.001) |
| Ln(BidSequence) | 0.003*** (0.0004) | 0.002*** (0.001) |
| Language | 0.026*** (0.0015) | 0.022*** (0.003) |
| Ln(Distance) | −0.002*** (0.0002) | −0.002*** (0.0003) |
| Observations | 560,529 | 526,754 |
| Project FEs | Yes | Yes |
| Service Provider FEs | No | Yes |
| Adj. $R^2$ | 0.049 | 0.068 |

*Notes*: Robust standard errors in parentheses; Observations reduced in Model 2 due to loss of singleton service providers.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

section, we explore the robustness of the results to our choice of estimator, the extent to which employers may have reduced their provision of worker feedback, as well as possible heterogeneity in the findings related to employers' experience with the dispute service.

### 5.2. Robustness Checks

We next considered the robustness of our results in a number of respects, beginning with our choice of estimator. Specifically, we assessed the bias and consistency of the LPM in our setting. Following the initial LPM estimation, we generated predicted values for the hiring outcome across our estimation sample. Horace and Oaxaca (2006) argue and demonstrate that bias is likely to be a problem with Linear Probability Models if a significant portion of these predictions falls outside the 0-1 interval. In our case, we observe that roughly 10% of the predictions fall outside the feasible range. We therefore consider a trimming estimator to assess the severity of the problem (Horrace and Oaxaca 2006). Specifically, we omit the observations associated with infeasible predictions and repeat the regression, which yields consistent results.

The results of this estimation are in Table 5, where we observe results consistent with those reported earlier (note that we omit coefficients associated with control variables which are identical in both sign and magnitude with the main findings, for the sake of brevity, similarly for the remaining estimations in the robustness check section). This finding provides some assurance that our results are not a result of bias or inconsistency of LPM estimators.

Next, we explored the robustness of our results to alternative measures. We began by considering an alternative measure of whether a project was *Objective*. Because our baseline measure of project objectivity derives from its listing category, there is the potential that our results are a product of some unobserved factor that systematically varies across categories. To address this possibility, we construct a second, more granular measure, which varies across projects (even within a category), based on human annotators from Amazon Mechanical Turk and text-based predictions. We first drew a random sample of 1000 projects from our dataset, which we posted to Amazon Mechanical Turk. We constructed a Human Intelligence Task (HIT) on MTurk, in which we instructed turkers to design a performance evaluation scheme for workers who would complete a given job, as described.

Job performance evaluation criteria and measures vary a great deal, and the literature has suggested that "different kinds of criterion measures will sometimes be required for different classes of jobs" (Gotfredson 1991). One distinction in the literature dealing with the design of job performance evaluation schemes, which bears particular relevance to our study, is that between people-oriented and things-oriented work. As Gotfredson (1991) states, "we would expect [supervisor] ratings to be used more often in people-oriented than things-oriented work." This literature

**Table 5 Effect of Ratings on Hiring Over Time by Project Type (LPM Trimming Estimator /w FEs)**

| Variable | (1) |
|---|---|
| Ln(Ratings) • $t-2$ | 0.001 (0.001) |
| Ln(Ratings) • $t-1$ | — |
| Ln(Ratings) • $t+0$ | −0.001 (0.001) |
| Ln(Ratings) • $t+1$ | 0.003 (0.001) |
| Ln(Ratings) • Objective • $t-2$ | 0.0003 (0.002) |
| Ln(Ratings) • Objective • $t-1$ | — |
| Ln(Ratings) • Objective • $t+0$ | −0.003* (0.0016) |
| Ln(Ratings) • Objective • $t+1$ | −0.004** (0.0016) |
| Observations | 467,547 |
| Project FEs | Yes |
| Service Provider FEs | Yes |
| Controls | Yes |
| Adj. $R^2$ | 0.0067 |

*Notes*: Robust standard errors in parentheses; Observations that hold a predicted value outside the feasible range following initial estimation (i.e., $y < 0 | y > 1$) omitted.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

points to two associated, broad classes of performance evaluation criteria: subjective and objective measures of job performance (Bommer et al. 1995). In line with that literature, we tasked crowd workers (turkers) with coding the relative appropriateness of either type of criterion for the projects (jobs) described in our data. Specifically, turkers were asked to read a job description and then recommend weightings between to two sources of information for the design of a given job's performance evaluation: subjective (i.e., the opinion of the person who wrote the job posting and who would make use of the resulting work output) or objective (e.g., comparing a worker's data entry results against a known accurate sample). The two weights summed to 100%.

Turkers provided their response using a slider that allowed them to weight results more toward manager opinion vs. automated metric-based evaluations. To ensure label quality, only Amazon Turk Masters were allowed to participate (turkers with an historical approval rating of at least 95%). Five turkers were hired to code each project. Turkers were paid \$0.15 per response. A total of 41 unique turkers completed the set of 5000 HITs ($5 \times 1000$ projects). Accordingly, we have many repeated observations per worker, which enables us to de-bias the responses with respect to individual worker perceptions/characteristics, via a worker fixed effect. We estimated a simple fixed effect regression, relating the resulting weighting measure (specifically, the percentage weight assigned to automatic/metric-based information) to vectors of project and worker fixed effects (dummies). Coefficients associated with the project dummies then served as our "de-biased" measure of work-task objectivity.

We then constructed a predictive model that relates the project description text to these objectivity labels. We built a corpus of all 100,000 + project descriptions. We pre-processed the corpus to i) remove all non-alpha-numeric characters, for example, punctuation, HTML tags, URLs, and stop words, ii) convert all text to lower case, and iii) apply stemming. We then extracted all unigrams, bigrams, and trigrams, constructing a term-document matrix (TDM) in which each cell contained the frequency count of each resulting token in each document. We subsequently removed any sparse tokens, namely those that did not appear at least 10 times in general, and within at least 10 project descriptions. Finally, we applied an inverse-document frequency (IDF) weighting. Using the resulting scaled TDM, we trained a random forest model that sought to the objectivity weighting label in the annotated sample, using the associated token frequencies. We implemented this model using the ranger package in R, a computationally efficient implementation of random forest that is well suited to dealing with high-dimensional data, as we have here (Wright and Ziegler 2017).

Although the textual data are likely to embed a great deal of noise, the predictive model performs reasonably well. As a naïve benchmark, we considered uniform predictions based on the mean of our ground-truth label (a value of 52% weighting assigned to "objective" inputs). Considering mean absolute error (MAE), the naïve prediction deviates from the hand-coded objectivity weight by 15 points in the labeled (training) sample. In contrast, the random forest model differs by just 6.8 points. Similarly, the naïve predictor exhibits a root mean squared error (RMSE) of 18.62, whereas the random forest model yields an RMSE of 9.12. The predictions are undoubtedly noisy, but this is unsurprising given the labeling itself is a subjective judgment. Nonetheless, the predictive model can recover some useful/relevant information about the validity of applying objective vs. subjective job performance criteria to our projects.

When we compare the distributions of predicted label values between projects falling into our initial groupings, that is, those derived from the project posting categories, we see a clear separation (Figure 3). This observation speaks to the face validity of both measures. Moreover, a $t$-test of mean differences in the MTurk imputed measure, between our objective and subjective project groupings derived from our category-based measure, yields a $t$-statistic of 150 and an associated $p$-value $< 0.00001$.

The continuous measure, derived using Mechanical Turk and text-based predictions, is noisy, in part because the labeling task is itself reasonably subjective, but also because prediction errors layer additional noise into the value. To reduce the influence of such noise in our regression analysis, we considered two remedies: we dichotomized the project-level measure, based on a median split, and we omitted observations associated with projects situated right at the mid-point of the AMT-measure's distribution. This approach of removing data that do not contribute to the prediction is common in prediction and classification tasks (e.g., Datta et al. 2006, Marchesotti et al. 2011). In practical terms, we removed projects where the AMT-derived measure of objectivity ranged between 50 and 53, that is, those closest to the median value of the continuous label. Employing the resulting binary indicator with our original regression specification, that is, replacing the category-based indicator of objectivity with our AMT-derived indicator (TurkObjective), we obtained the results reported in Table 6, which are consistent with those obtained using our category-based measure. Repeating this analysis using the raw, continuous AMT-derived label, we observe a consistent pattern of results. We prefer this binary measure because it is more readily

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

**Figure 3 Distribution of Projects' AMT Objectivity Values by Category-based Measure**

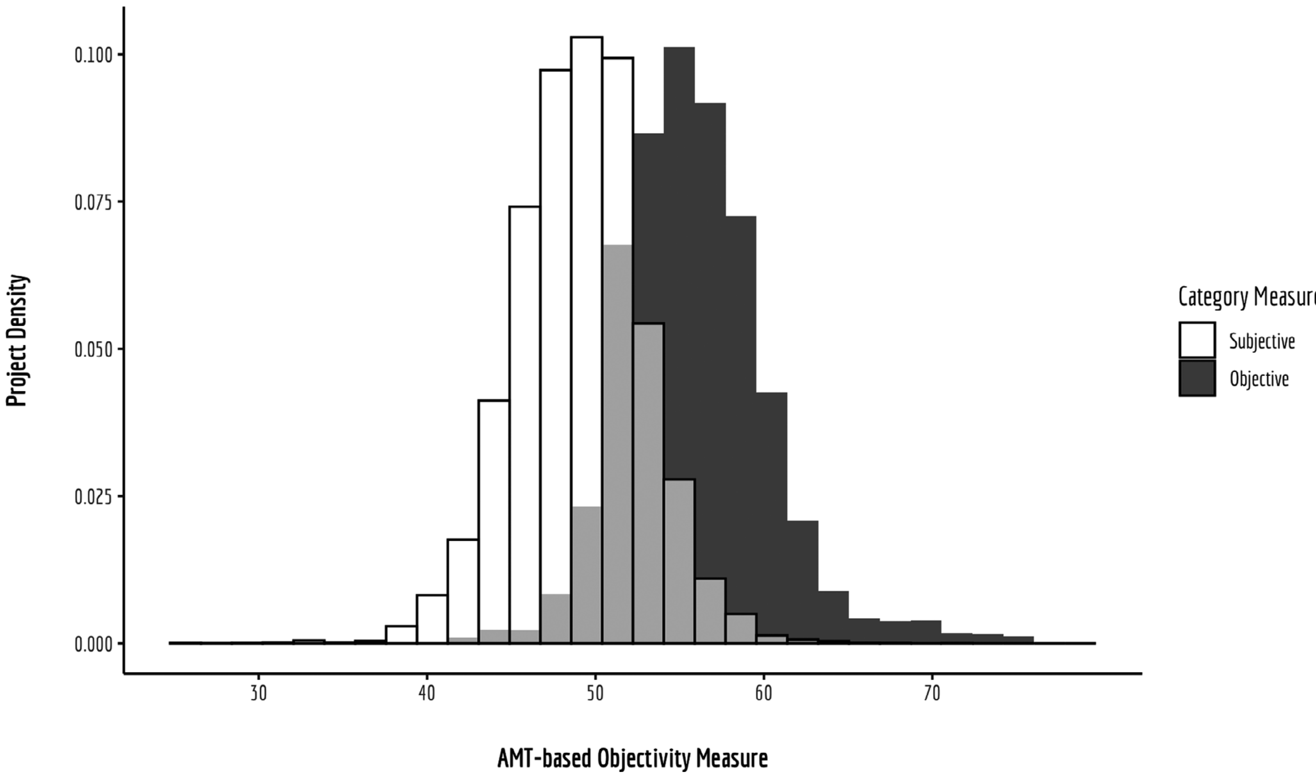



**Table 6 Effect of Ratings on Hiring Over Time by Project Type (LPM / w Fes—Based on Amazon Mechanical Turk Rating)**

| Variable | (1) |
|---|---|
| Ln(Ratings) • $t-2$ | 0.001 (0.001) |
| Ln(Ratings) • $t-1$ | — |
| Ln(Ratings) • $t+0$ | −0.001 (0.001) |
| Ln(Ratings) • $t+1$ | −0.001 (0.001) |
| Ln(Ratings) • TurkObjective • $t-2$ | 0.0005 (0.002) |
| Ln(Ratings) • TurkObjective • $t-1$ | — |
| Ln(Ratings) • TurkObjective • $t+0$ | $-0.003^{+}$ (0.0016) |
| Ln(Ratings) • TurkObjective • $t+1$ | $-0.003^{+}$ (0.0016) |
| Observations | 381,639 |
| Project FEs | Yes |
| Service Provider FEs | Yes |
| Controls | Yes |
| Adj. $R^2$ | 0.0059 |

*Notes*: Standard errors in parentheses; Observations omitted for projects at the midpoint of AMT-derived objective-subjective label distribution.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

interpretable and comparable to the objectivity measure employed in our other analyses.

Next, we considered an alternative measure that can serve as an ex-ante signal of worker quality. Specifically, we tried replacing our measure of *Ratings* with an indicator capturing whether a service provider possesses a platform-provided skill certification (*HasCert*). Repeating our main regression analysis, we obtained consistent results; as can be seen in Table 7, we observed the same pattern of effects. Note that, as before, we omit coefficients associated with control variables for the sake of brevity.

Third, we tried re-estimating our models across the different project work categories. That is, we split our sample into individual project categories (Data Entry vs. IT & Software vs. Writing vs. Logo Design) and repeat our analysis. These subsample regressions allow us to ensure that we do not observe contradictory estimates in any of the individual product categories, that might run counter to the overall notion of how we predict rating influence will shift between *Objective* and *Subjective* project categories. We report our four sub-sample regressions below in Table 8. The results confirm that the estimates behave as expected within each project category.

### 5.3. Additional Analyses

Having established the robustness of our results, we next turn to some empirical extensions. First, we consider whether the observed changes in employer hiring behavior were beneficial or detrimental for project outcomes. The results documented by Pallais (2014) on Upwork—that buyers' tendency to under-value service providers who lack reputation is generally to their detriment—suggests that a shift away from relying on ratings in hiring decisions should not be detrimental, on average. It may be, for example, that the dispute service, offered in tandem with a reputation system, would be a more effective deterrent for

**Table 7 Effect of Certifications on Hiring Over Time by Project Objectivity (LPM /w FEs)**

| Variable | (1) |
|---|---|
| HasCert • $t-2$ | −0.003 (0.004) |
| HasCert • $t-1$ | — |
| HasCert • $t+0$ | −0.002 (0.003) |
| HasCert • $t+1$ | 0.002 (0.003) |
| HasCert • Objective • $t-2$ | 0.001 (0.006) |
| HasCert • Objective • $t-1$ | — |
| HasCert • Objective • $t+0$ | −0.012* (0.005) |
| HasCert • Objective • $t+1$ | −0.014** (0.005) |
| Observations | 526,754 |
| Project FEs | Yes |
| Service Provider FEs | Yes |
| Controls | Yes |
| Adj. $R^2$ | 0.068 |

*Notes*: Robust standard errors in parentheses; Observations reduced in Model 2 due to loss of singleton service providers.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

opportunism. Although we cannot isolate the direct effect of the dispute service on project outcomes from the indirect effect (via changes in the hiring calculus), we can assess the net relationship. Notably, examining this also enables us to evaluate whether the introduction of the dispute resolution system reduced the informativeness of the reputation system by reducing buyers' tendency to enter new feedback. We undertake this analysis by estimating equation (2).

$$\begin{aligned} EPP_{w,e} &= \lambda_1 \cdot \sum_t \tau_{w,e} + \lambda_2 \cdot Objective_{w,e} \\ &\quad + \lambda_3 \cdot Objective_{w,e} \cdot \sum_t \tau_p + C_{w,e} + \alpha_w + \delta_e + \varepsilon_{w,e}. \end{aligned} \tag{2}$$

Our outcome of interest is a proxy for buyer satisfaction with the work, a binary indicator of effective positive feedback, or EPP, based on the measure proposed by Nosko and Tadelis (2015). As noted earlier, the valence of reviews is not particularly informative in our setting because 95% of service providers have an average rating above 9/10. However, the absence of a rating for a transaction is often telling in this environment—Nosko and Tadelis (2015) refer to such unrated transactions as "silent" expressions of satisfaction.

These silent transactions may occur because a transaction partner exits the platform without leaving feedback after a bad experience or, in some settings, because individuals are wary of reciprocal feedback. Nosko and Tadelis (2015) propose an alternative reputation valence that captures silent transactions and demonstrate its efficacy in the context of eBay. We take the same general approach here and model our binary indicator of whether a buyer decides to leave a positive rating (9/10 or 10/10).

We estimate this regression with a reduced sample of service provider-project pairs, indexed by $w$ and $e$, respectively, namely those pairs involving service providers who were awarded a contract by the seller. We regress the outcome on our time dummies and an interaction with the *Objective* indicator. Once again, we condition on a set of variables, $C$, that vary across service provider–seller observations, including geographic distance, the contracted dollar amount, shared language, and so on. Our coefficients of interest are $\lambda_1$ and $\lambda_2$, with the former reflecting changes over time in the probability of buyer satisfaction and feedback provision, and the latter reflecting any relative differences in the trend of buyer satisfaction and feedback provision for *Objective* work. Because this regression considers only one observation per project, we are unable to incorporate a project fixed effect. However, we can include a buyer fixed effect, $\delta$, because we observe the same buyer across multiple projects. Note that, because we omit the project fixed effect, they no longer subsume the main effects of our time dummies, $\tau$.

We present the results of these analyses in Table 9. We observe that the introduction of the dispute

**Table 8 Category-Specific Rating Effects Over Time (LPM /w FEs)**

| | <== | ***More Objective*** | ***More Subjective*** | ==> |
|---|---|---|---|---|
| Variable | (1) Data Entry | (2) IT & Software | (3) Writing | (4) Design |
| Ln(Ratings) • $t-2$ | −0.003 (0.003) | −0.002 (0.003) | 0.001 (0.002) | 0.0017 (0.001) |
| Ln(Ratings) • $t-1$ | — | — | — | — |
| Ln(Ratings) • $t+0$ | −0.003** (0.001) | −0.009*** (0.003) | −0.001 (0.002) | −0.001 (0.001) |
| Ln(Ratings) • $t+1$ | −0.005*** (0.002) | −0.003 (0.003) | −0.001 (0.002) | −0.0004 (0.001) |
| Observations | 149,338 | 84,489 | 94,464 | 232,297 |
| Project FEs | Yes | Yes | Yes | Yes |
| Service Provider FEs | Yes | Yes | Yes | Yes |
| Controls | Yes | Yes | Yes | Yes |
| Adj. $R^2$ | 0.071 | 0.056 | 0.081 | 0.051 |

*Notes*: Robust standard errors reported in parentheses.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

system led to a systematic increase in employers' provision of positive feedback. Once again, we see that this occurs, in particular, for *Objective* project work. Accordingly, we conclude that our main results are not driven by a decline in the provision of ratings, and moreover that the shift away from rating reliance was apparently not detrimental for project outcomes. To the contrary, our findings suggest an increase in matching quality. This, in turn, suggests that dispute resolution systems can induce more efficient matching outcomes. While we do not empirically analyze the mechanisms underlying this effect, we surmise that the better performance is likely driven by the mechanisms identified by Pallais (2014), or alternatively by inducing service providers to be more thoughtful about the jobs they pursue (e.g., self-selection into jobs they are truly capable of performing well).

Finally, we explored additional moderators that capture employers' experience with the dispute resolution system. In particular, we constructed two binary indicators, based on actual dispute records, which respectively capture whether, as of a project submission date, the buyer had previously ever had a dispute filed against them, or had filed a dispute against another party. That is, we consider the possibility that the shift in rating consideration depends on whether employers perceive the service to be effective in achieving its objective of inducing good behavior. We perform this analysis by introducing two new binary moderators, which respectively capture whether an employer has either i) previously filed a complaint against a service provider, or ii) previously had a complaint filed against them. The results of these analyses are in Table 10 and Table 11.

**Table 9 Effect of Dispute Service Introduction on Buyers' Provision of Feedback (LPM /w FEs)**

| Variable | (1) |
|---|---|
| Ln(Ratings) • $t-2$ | −0.004 (0.019) |
| Ln(Ratings) • $t-1$ | — |
| Ln(Ratings) • $t+0$ | −0.006 (0.017) |
| Ln(Ratings) • $t+1$ | −0.038 (0.020) |
| Ln(Ratings) • Objective • $t-2$ | 0.026 (0.028) |
| Ln(Ratings) • Objective • $t-1$ | — |
| Ln(Ratings) • Objective • $t+0$ | 0.024 (0.024) |
| Ln(Ratings) • Objective • $t+1$ | 0.065** (0.026) |
| Observations | 31,025 |
| Project FEs | Yes |
| Service Provider FEs | Yes |
| Controls | Yes |
| Adj. $R^2$ | 0.325 |

*Notes*: Robust standard errors in parentheses; Observations omitted for projects at the midpoint of AMT-derived objective–subjective label distribution; Observations reduced in Model 2 due to loss of singleton service providers.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

**Table 10 Effect of Ratings on Hiring Over Time /w Moderator for Complaint Filing (LPM /w FEs)**

| Variable | (1) |
|---|---|
| Ln(Ratings) • $t-2$ | 0.0014+ (0.0007) |
| Ln(Ratings) • $t-1$ | — |
| Ln(Ratings) • $t+0$ | −0.002** (0.001) |
| Ln(Ratings) • $t+1$ | −0.0018** (0.0007) |
| Ln(Ratings) • $t-2$ • Has Made Complaint | — |
| Ln(Ratings) • $t-1$ • Has Made Complaint | — |
| Ln(Ratings) • $t+0$ • Has Made Complaint | 0.011 (0.032) |
| Ln(Ratings) • $t+1$ • Has Made Complaint | 0.015 (0.032) |
| Observations | 526,754 |
| Project FEs | Yes |
| Service Provider FEs | Yes |
| Adj. $R^2$ | 0.068 |

*Notes*: Robust standard errors in parentheses; We omit controls and interactions terms involving the dispute system experience moderator for the sake of brevity.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

We observe that, although buyers who have previously filed a complaint display the same decline in rating consideration after the dispute system appears, this is less true of employers who receive a complaint. That is, buyers who receive a complaint continue to pay attention to service provider ratings in their hiring decisions. While suggestive, this result indicates that the impact of dispute system introduction on the role of ratings in hiring decisions does indeed depend on buyers' experiences with the dispute process. Buyers who suffer as a result of the dispute system rely upon it less.

## 6. General Discussion

We have reported a set of empirical analyses of a natural experiment in which a dispute resolution

**Table 11 Effect of Ratings on Hiring Over Time /w Moderator for Complaint Receipt (LPM /w FEs)**

| Variable | (1) |
|---|---|
| Ln(Ratings) • $t-2$ | 0.0014+ (0.0007) |
| Ln(Ratings) • $t-1$ | — |
| Ln(Ratings) • $t+0$ | −0.002** (0.001) |
| Ln(Ratings) • $t+1$ | −0.0017* (0.0007) |
| Ln(Ratings) • $t-2$ • Has Received Complaint | — |
| Ln(Ratings) • $t-1$ • Has Received Complaint | — |
| Ln(Ratings) • $t+0$ • Has Received Complaint | 0.058* (0.025) |
| Ln(Ratings) • $t+1$ • Has Received Complaint | 0.055* (0.025) |
| Observations | 526,754 |
| Project FEs | Yes |
| Service Provider FEs | Yes |
| Adj. $R^2$ | 0.068 |

*Note*: Robust standard errors in parentheses; We omit controls and interactions terms involving the dispute system experience moderator for the sake of brevity.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.10$.

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

system was introduced into a leading online service procurement platform, to understand whether and to what extent this form of ex-post, platform-provided arbitration can address supplier quality information asymmetries, substituting for buyers' reliance upon common informational or signaling mechanisms, such as online reputation systems or platform-provided skill certifications. We show that reputation and certifications have a weaker influence on hiring decisions in the presence of a dispute resolution system, particularly for "objective" work categories.

Most broadly, our work speaks directly to recent calls in the OM literature for an improved understanding of mechanism design with an eye toward i) mitigating information asymmetry, and ii) improving matching efficiency (Chen et al. 2020, Hu 2020). Our work provides a novel consideration of the value of an ex-post mechanism for fostering positive transaction outcomes. We provide evidence that dispute resolution systems can serve as an effective substitute to ex-ante information mechanisms, like reputation systems and skill screening certifications, for some types of projects.

Our findings indicate that dispute services can help reduce the cold-start problem for some types of work and, in turn, can reduce barriers to entry for new service providers in the market, which past work suggests will improve the efficiency of the market (Pallais 2014). Indeed, we observe that buyers are significantly more likely to provide favorable ratings of service providers after they gain access to the dispute system. We surmise this may occur because of a reduction in opportunism on the part of service providers, following transaction, or it may arise because service providers' knowledge of the presence of the dispute system may induce them to be more thoughtful or selective in the jobs they apply for.

Importantly, however, our findings appear to only apply to work where objective performance evaluation is possible. This finding suggests a further qualification to the conclusions from the existing operations management literature, such as those from Allon et al. (2017). We observe that the dispute resolution system reduces reliance upon certifications in hiring as well, although once again this is contingent on the work being objective in nature. The particular finding that certifications play a weaker role is also essential to keep in mind from an operations management perspective, as many platform operators offer certification services for a fee. Our findings indicate that platform operators should remain cognizant that demand for these certifications may shift with the introduction of dispute resolution systems. Moreover, platform operators should perhaps focus on recasting their certification offerings, for example, placing greater focus on certifications in Writing and Design work.

Our work also empirically verifies some game-theoretic predictions of Bakos and Dellarocas (2011), in their comparison of the relative efficacy of litigation and reputation-based mechanisms for attracting high-quality sellers or service providers to a marketplace. First, according to their model, the most efficient mechanism depends on the degree to which the dispute resolution system is reliable in its outcomes, the cost of using the dispute service, and the value of expected damages. In our context, the fee associated with using the dispute resolution system is the higher of \$5 or 5%, with the fee refunded to the winning party. Additionally, damages awarded are limited to the value of the project. Given that the cost of using dispute resolution and the expected damages are in some sense homogeneous across projects, assuming roughly equivalent project budgets, the effects we observe are attributable to variation over job types in the expected reliability of the dispute resolution outcome. Thus, when it comes to well codified, objectively evaluable services and tasks, such as data entry, the impact of reputation decreases with the introduction of the formal dispute resolution system.

Our work is subject to several limitations. First, our analyses of rating effects focus solely on rating volumes, ignoring valence. As noted earlier, this is because of rating inflation in the marketplace, such that the vast majority of ratings that arrive in the market are extremely positive (95% of service providers hold an average rating above 9/10). As has been documented in the prior literature on online reviews and ratings, rating inflation is a persistent problem in many markets, for example, because users are concerned about reciprocity on the part of their transaction partner, should they enter a negative review (Zervas et al. 2015). Other recent work has documented other systematic processes by which ratings inflate over time (Filippas et al. 2019). However, some recent work has also proposed alternative implementations of rating systems that can help to mitigate inflation, for example, by taking advantage of information implied by "silent'"(unrated) transactions (Nosko and Tadelis 2015) or precluding retaliatory behavior (Ye et al. 2014). As such, valence information may play an essential role in other markets, which may also provide an interesting source of variance to explore, as dispute services may have asymmetric effects on hiring outcomes for positively vs. negatively rated service providers, for example, perhaps buyers will be more likely to take a chance on a negatively rated service provider. Second, our data and analyses do not enable us to examine the cost born by the platform to operate the dispute service. In part, this is because we do not have precise costs associated with managing a dispute. More generally, however, this is because many aspects of the cost are intangible. A buyer who is unhappy with a dispute outcome may leave the platform never to return, as might a service provider. These sorts of unobservable factors make it difficult to properly assess the welfare implications of a dispute resolution system empirically.

There is ample opportunity for future work to build on our findings. First, future work can look to empirically validate other aspects of game-theoretic predictions around the role of dispute services vs. reputation systems. For example, our context does not enable us to examine the influence of variation in fees associated with initiating disputes, or differences in damages awarded, primarily because our natural experiment does not provide a source of variance around these factors. Second, future work might delve more deeply into the dynamics of dispute system use and reliance on reputation. Our analysis focuses merely on the average effect across buyers, in the presence of a newly introduced dispute service. However, buyer behavior is quite likely to shift as a function of the buyer's experience with the service, conditional on use. If a buyer experiences a favorable outcome in a dispute, he or she is perhaps more likely to rely on the service going forward. Conversely, buyers who lose a dispute may be less likely to rely upon it. These effects have the potential to operate independently of project characteristics, namely subjectivity or objectivity of quality evaluations.

Additionally, the services that are characterized more by subjective fit are also, notably, those likely to necessitate greater collaboration between buyers and service providers. Accordingly, a possible alternative mechanism that may underlie our finding of heterogeneity is perhaps to do with this collaborative nature of creative work, which prior work has suggested (Sun and Xu 2018). Given that creative tasks, being subjective, are more likely to necessitate collaboration between the buyer and the service provider, dispute services may have relatively weaker effects with creative work because the separation of fault between the buyer and the service provider is altogether "fuzzier" or more ambiguous. Future work is needed to assess this possibility.

Another interesting possibility for future work is related to what is suggested by Anderson et al. (2009). Those authors present a method for assessing the value of ex-post product return options, across different buyers and product types. A potential avenue for future work might be to explore the option value of dispute resolution systems across both buyers and sellers in online service marketplaces, as well as across project types. Such work could inform platform decisions about how to design dispute policies, as well as whether and where to institute dispute policies.

## 7. Conclusion

The on-demand digital service platforms have recently started to gain significant attention from operations management and information systems researchers (Allon et al. 2017, Hong and Pavlou 2017, Taylor 2018). Given the semi-anonymous nature, one important issue that often arises in such platforms is the disagreements and disputes among the buyers and service providers. This work offers what is, to our knowledge, the first empirical consideration of the interaction between online reputation systems and a platform-provided dispute resolution system. Extending prior work that has only examined the determinants of dispute initiation, or the dynamics of the dispute process itself, and the associations with the dispute outcome, we empirically examine how participant behavior in the service procurement platforms may shift in the presence of a dispute resolution system. Our findings have direct policy implications for platform operators, as they suggest potentially efficiency gains by overcoming the reputation cold-start problem, yet simultaneous concerns associated with the weakened value of platform-provided certifications. That said, more research is needed to understand this domain. Open questions remain around the antecedents of dispute service introduction by platform operators, the costs of these services to the platform provider, and the optimal design and delivery of dispute services. We hope that this work can offer a first foray into this line of questioning, stimulating subsequent work on the topic.

## Acknowledgments

The authors are grateful to Paul Pavlou and Nina Huang for helpful early discussions on this work. Gordon Burtch acknowledges funding from the National Science Foundation [Award SCC / 1952085].

### Notes

[1]http://www.fastcompany.com/3005402/ebay-spinoff-modria-judge-judy-cyber-shoppers
[2]www.getafreelancer.com.
[3]We estimate this regression employing Stata's *reghdfe* procedure, which allows efficient estimation of multiple high-dimensional fixed effects models (Correia 2018).
[4]Project fixed effects subsume the direct influence of time fixed effects, given that projects are associated with the time of their posting; however, we remain capable of estimating the moderating influence of time, as this amounts to an assessment of how the moderated factor's influence varies between earlier and later project postings.

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

## References

Allon, G., A. Bassamboo, E. B. Çil. 2017. Skill management in large-scale service marketplaces. *Prod. Oper. Manag.* **26**(11): 2050–2070.

Anderson, E. T., K. Hansen, D. Simester. 2009. The option value of returns: Theory and empirical evidence. *Marketing Science* **28**(3): 405–423.

Autor, D. H., D. Dorn. 2013. The growth of low-skill service jobs and the polarization of the US labor market. *Am. Econ. Rev.* **103**(5): 1553–1597.

Autor, D. H., F. Levy, R. J. Murnane. 2003. The skill content of recent technological change: An empirical exploration. *Q. J. Econ.* **118**(4): 1279–1333.

Ba, S., P. A. Pavlou. 2002. Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly* **26**(3): 243–268.

Bakos, Y., C. Dellarocas. 2011. Cooperation without enforcement? A comparative analysis of litigation and online reputation as quality assurance mechanisms. *Management Sci.* **57**(11): 1944–1962.

Banker, R., I. Hwang. 2008. Importance of measures of past performance: Empirical evidence on quality of e-service providers. *Contemporary Account. Res.* **25**(2): 307–337.

Benjaafar, S., M. Hu. 2020. Operations management in the age of the sharing economy: What is old and what is new? *Manuf. Serv. Oper. Manag.* **22**(1): 93–101.

Bommer, W. H., J. L. Johnson, G. A. Rich, P. M. Podsakoff, S. B. MacKenzie. 1995. On the interchangeability of objective and subjective measures of employee performance: A meta-analysis. *Pers. Psychol.* **48**(3): 587–605.

Brett, J. M., M. Olekalns, R. Friedman, N. Goates, C. Anderson, C. Lisco. 2007. Sticks and stones: Language, face, and online dispute resolution. *Acad. Manage. J.* **50**(1): 85–99.

Chen, Y. J., T. Dai, C. G. Korpeoglu, E. Körpeoğlu, O. Sahin, C. S. Tang, S. Xiao. 2020. OM forum—innovative online platforms: Research opportunities. *Manuf. Serv. Oper. Manag.* **22**(3): 430–445.

Correia, S. 2018. *REGHDFE: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects*. Available at: https://EconPapers.repec.org/RePEc:boc:bocode:s457874. Accessed Nov. 23rd, 2020.

Culnan, M. J., P. K. Armstrong. 1999. Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organ. Sci.* **10**(1): 104–115.

Datta, R., D. Joshi, J. Li, J. Z. Wang 2006. Studying aesthetics in photographic images using a computational approach. In European conference on computer vision. Springer, Berlin, Heidelberg, 288–301.

Dellarocas, C. 2003. The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Sci.* **49**(10): 1407–1424.

Einav, L., C. Farronato, J. Levin. 2016. Peer-to-peer markets. *Annual Rev. Econ.* **8**(1): 615–635.

Filippas, A., J. J. Horton, J. M. Golden. 2019. Reputation Inflation. National Bureau of Economic Research. no. w25857. Available at: https://ssrn.com/abstract=3391010 (Accessed Nov. 23rd, 2020).

Fradkin, A., E. Grewal, D. Holtz. 2018. The determinants of online review informativeness: Evidence from field experiments on Airbnb. SSRN Working Paper. Available at: https://ssrn.com/abstract=2939064 (Accessed Nov 23rd, 2020).

Friedman, R., C. Anderson, J. Brett, M. Olekalns, N. Goates, C. Lisco. 2004. The positive and negative effects of anger on dispute resolution: Evidence from electronically mediated disputes. *J. Appl. Psychol.* **89**(2): 369–376.

Fugger, N., E. Katok, A. Wambach. 2019. Trust in procurement interactions. *Management Sci.* **65**(11): 4951–5448.

Garg, N., R. Johari. 2018. Designing informative rating systems for online platforms: Evidence from two experiments. Working Paper. Available at: https://deepai.org/publication/designing-informative-rating-systems-for-online-platforms-evidence-from-two-experiments (Accessed Nov. 23rd, 2020).

Gottfredson, L. S. 1991. The evaluation of alternative measures of job performance. A. K. Wigdor, Green Jr, B. F., eds. In *Performance assessment for the workplace*, Tech. Issues. vol. 2. Natl. Acad. Sci, Washington, DC, 75–126.

Hong, Y., P. A. Pavlou. 2017. On buyer selection of service providers in online outsourcing platforms for IT services. *Inform. Systems Res.* **28**(3): 547–562.

Hong, Y., B. B. M. Shao 2020. On Factors That Moderate The Effect Of Buyer-Supplier Experience on E-procurement platforms. *Prod. Oper. Manag.* https://doi.org/10.1111/poms.13291

Hong, Y., C. Wang, P. A. Pavlou. 2016. Comparing open and sealed bid auctions: Evidence from online labor markets. *Inform. Systems Res.* **27**(1): 49–69.

Horrace, W. C., R. L. Oaxaca. 2006. Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Econ. Lett.* **90**(3):321–327.

Horton, J. J. 2019. Buyer uncertainty about seller capacity: Causes, consequences, and a partial solution. *Manage. Sci.* **65**(8): 3518–3540.

Hotelling, H. 1929. Stability in competition. *Econ. J.* **39**(153): 41–57.

Hu, M. 2020. From the classics to new tunes: A neoclassical view on sharing economy and innovative marketplaces. *Prod. Oper. Manag.* (special issue of Frontier Analytic Modeling and Methods for Operations Management), Forthcoming. https://doi.org/10.1111/poms.13330

Hu, N., P. A. Pavlou, J. J. Zhang. 2017. On self-selection biases in online product reviews. *MIS Quarterly* **41**(2): 449–471.

Hu, N., J. J. Zhang, P. A. Pavlou. 2009. Overcoming the j-shaped distribution of product reviews. *Commun. ACM* **52**(10): 144–147.

Kokkodis, M. 2019. Reputation deflation through dynamic expertise assessment in online labor markets. WWW '19: The World Wide Web Conference. 896–905 (May).

Kokkodis, M., P. G. Ipeirotis. 2015. Reputation transferability in online labor markets. *Management Sci.* **62**(6): 1687–1706.

Lim, W. S. 2001. Producer-supplier contracts with incomplete information. *Management Sci.* **47**(5): 709–715.

Lin, M., Y. Liu, S. Viswanathan. 2018. Effectiveness of reputation in contracting for customized production: Evidence from online labor markets. *Management Sci.* **64**(1): 345–359.

Macinnes, I., Y. Li, W. Yurcik. 2005. Reputation and dispute in eBay transactions. *Int. J. Electron. Commerce* **10**(1): 27–54.

Marchesotti, L., F. Perronnin, D. Larlus, G. Csurka 2011. Assessing the aesthetic quality of photographs using generic image descriptors. International Conference on Computer Vision (pp. 1784–1791). IEEE.

Moreno, A., C. Terwiesch. 2014. Doing business with strangers: Reputation in online service marketplaces. *Inform. Systems Res.* **25**(4): 865–886.

Nosko, C., S. Tadelis. 2015. The limits of reputation in platform markets: An empirical analysis and field experiment. *Natl Bur. Econ. Res.* Available at: https://ssrn.com/abstract=2548349 (Accessed Nov. 23rd, 2020).

Pallais, A. 2014. Inefficient hiring in entry-level labor markets. *Am. Econ. Rev.* **104**(11): 3565–3599.

Pavlou, P. A., D. Gefen. 2004. Building effective online marketplaces with institution-based trust. *Inform. Sys. Res.* **15**(1): 37–59.

Snir, E. M., L. M. Hitt. 2003. Costly bidding in online markets for IT services. *Management Sci.* **49**(11): 1504–1520.

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341

Sun, H., L. Xu. 2018. Online reviews and collaborative service provision: A signal-jamming model. *Prod. Oper. Manag.* **27**(11): 1960–1977.

Swinney, R., S. Netessine. 2009. Long-term contracts under the threat of supplier default. *Manuf. Serv. Oper. Manag.* **11**(1): 109–127.

Tadelis, S. 2016. Reputation and feedback systems in online platform markets. *Annual Rev. Econ.* **8**(1): 321–340.

Tang, C. S. 2006. Perspectives in supply chain risk management. *Int. J. Prod. Econ.* **103**(2): 451–488.

Taylor, T. A. 2018. On-demand service platforms. *Manuf. Serv. Oper. Manag.* **20**(4): 704–720.

Tunca, T. I., D. Wu, F. Zhong. 2014. An empirical analysis of price, quality, and incumbency in procurement auctions. *Manuf. Serv. Oper. Manag.* **16**(3): 346–364.

Wang, Y., J. Yang, L. Qi. 2017. A game-theoretic model for the role of reputation feedback systems in peer-to-peer commerce. *Int. J. Prod. Econ.* **191**(C): 178–193.

Wright, M. N., A. Ziegler. 2017. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* **77**(1): 1–17.

Ye, S., G. G. Gao, S. Viswanathan. 2014. Strategic behavior in online reputation systems: Evidence from revoking on eBay. *MIS Quarterly* **38**(4): 1033–1056.

Zervas, G., D. Proserpio, J. Byers. 2015. A first look at online reputation on Airbnb, where every stay is above average. SSRN Working Paper. Available at: https://ssrn.com/abstract=2554500 (Accessed Nov. 23rd, 2020).

Please Cite this article in press as: Burtch, G., Y., Hong, et al. When Does Dispute Resolution Substitute for a Reputation System? Empirical Evidence from a Service Procurement Platform. *Production and Operations Management* (2021), https://doi.org/10.1111/poms.13341