

Reputation Inflation*

Apostolos Filippas
Fordham

John J. Horton
MIT & NBER

Joseph M. Golden
Collage.com

December 4, 2020

Abstract

A solution to marketplace information asymmetries is to have trading partners publicly rate each other post-transaction. Many have shown that these ratings are effective; we show that their effectiveness can deteriorate over time. The problem is that ratings are prone to inflation, with raters feeling pressure to leave “above average” ratings, which in turn pushes the average higher. This pressure stems from raters’ desire to not harm the rated seller. As the potential to harm is what makes ratings effective, reputation systems, as typically designed, become less informative in the long-run.

*Thanks to David Holtz, Ramesh Johari, Nico Lacetra, Xiao Ma, Foster Provost, Aaron Sojourner, and Richard Zeckhauser for very helpful comments and suggestions. Author contact information and code are currently or will be available at <http://www.john-joseph-horton.com/>.

1 Introduction

Scores of various kinds—credit scores, school grades, restaurant and film “star” reviews, restaurant hygiene scores, Better Business Bureau ratings—have long been important sources of information for market participants. A large literature documents the economic importance of such scores (Resnick et al., 2000; Jin and Leslie, 2003; Resnick et al., 2006; Mayzlin et al., 2014; Ghose et al., 2014; Luca, 2016; Luca and Zervas, 2016). As more of economic and social life has become computer-mediated, opportunities to generate and apply new kinds of scores—particularly in marketplace contexts—have proliferated, as has the number of individuals and businesses subject to these “reputation systems” (Levin, 2011; Hall and Krueger, 2018; Katz and Krueger, 2019). Designing effective reputation systems has become a first-order question in the digital economy.

In online marketplaces, “reputations” are typically calculated from numerical feedback scores left by past trading partners. As many have noted, the distribution of feedback scores in various online marketplaces seems implausibly rosy.¹ For example, the median seller on eBay has a score of 100% positive feedback ratings, and the tenth percentile is 98.21% positive feedback ratings (Nosko and Tadelis, 2015). On Uber and Lyft, it is widely known that anything less than 5 stars is considered “bad” feedback: Athey et al. (2019) find that nearly 90% of UberX Chicago trips in early 2017 had a perfect 5-star rating.

Of course, there is no ground truth that tells us what the distribution of scores in some market should look like at a moment in time. However, if we look at how the distribution of scores changes over time, we can potentially learn more about what the reputation system is measuring. Two distinct—but not mutually exclusive—reasons can cause rising feedback scores: (1) raters are becoming more satisfied, or (2) raters are lowering their rating standards. The first possibility—more satisfied raters giving higher scores—is due to improvements in market “fundamentals,” such as better marketplace features, better cohorts of buyers/sellers joining the platform, low-quality buyers/sellers exiting the platform, and lower-priced products. The second possibility—giving higher scores despite not being more satisfied—can be thought of as a kind of inflation. This “reputation inflation” erodes the comparability of feedback scores over time, and can reduce the informativeness of a reputation system.²

In this paper, we examine the reputation system of a large online labor market, focusing on the evolution of average feedback scores over time, and the causes for the dynamics we observe. Mirroring findings from other marketplaces, we find that the distribution of

¹A PEW research center survey finds that while more than 80% of U.S. adults read online reviews before purchasing an item, almost 50% find the truthfulness of these reviews hard to assess. For more details, see <http://www.pewinternet.org/2016/12/19/online-reviews>.

²Reputation inflation is similar to the conjecture about the increase in college grades—namely, that students’ work is not getting better, but rather that the same quality of work now earns a higher grade (Babcock, 2010; Butcher et al., 2014).

recent employer feedback for workers is highly top-censored, with an overwhelming majority receiving perfect feedback.³ However, the distribution has not always been this skewed—the fraction of workers receiving a perfect 5-star rating grew from 33% to 85% in just 6 years. Feedback scores in four other online marketplaces for which we obtained longitudinal data exhibit a similar increase over time.

Our first task is to examine how much of the observed increase in feedback scores cannot be attributed to higher rater satisfaction caused by improvements in fundamentals, but rather to raters lowering their rating standards. Toward that end, we use longitudinal data that include both the primary feedback score and an alternative measure of rater satisfaction. The idea is that if rater satisfaction is also captured by an alternative measure that does not inflate—or inflates at a lower rate—then we can exploit this difference to produce an estimate of inflation in the measure of interest. Our approach allows us to circumvent the challenges associated with estimating the effect of improvements in platform fundamentals on employers’ utilities.

As a first alternative measure of rater satisfaction, we use information obtained by the introduction of a parallel and experimental reputation system that asked employers to rate workers “privately.” This private feedback was not conveyed to the rated workers, nor made public to future would-be employers. At the same time, raters were still asked to give the status quo “public” feedback, both written and numerical. The conjecture motivating the platform’s private feedback feature was that employers would be more candid in private, willing to give bad feedback if not exposed to retaliation from angry workers, and/or because a bad report would not harm the worker. We find that average private feedback scores were decreasing over the period they were collected, but at the same time average public feedback scores for the *same* transactions were increasing. This difference helps to disentangle the two reasons that can cause rising feedback scores: employers were not becoming more satisfied, but instead they were lowering their rating standards for public feedback.

As a second alternative measure of rater satisfaction, we use the sentiment employers express in the written feedback that accompanies numerical scores. To capture this sentiment, we fit a model that predicts numerical feedback from the text of written feedback. Critically, the model is fit using written feedback from a narrow window of time early in our data, allowing us to learn the relationship that prevailed between written sentiment and numerical score when the training feedback text was created. Using the predictions of the model, we can then decompose the growth in average feedback scores into a component due to improvements in market fundamentals that increased rater satisfaction and is reflected in written feedback, and the residual component that is due to inflation. We find that although predicted feedback scores based on written feedback have increased over time, presumably due to improvements

³We use the terms “employer” and “worker” for consistency with the literature, and not as a comment on the legal relationship of the transacting parties.

in fundamentals, they have not increased nearly as much as the numerical feedback scores. Our estimates suggest that more than 50% of the increase in scores over a 6 year period was due to inflation, and are robust across different specifications and training sets. Furthermore, insofar written feedback is also subject to inflation, our approach *understates* the extent of reputation inflation.

As a less model-dependent approach, we also compare the numerical feedback scores associated with the same common sentences appearing in written feedback in two different time periods. We show that the same sentences systematically have higher associated numerical feedback scores in the latter period. For example, employers calling the work they received “terrible” would assign on average a public feedback score of 1.4 stars in 2008, but they would instead assign 2.4 stars in 2015.⁴

We next turn to understanding the cause of reputation inflation. We argue that the key to understanding reputation inflation is appreciating the role of costs, and how these costs affect raters. A starting point is the divergence between public and private feedback scores: 28.4% of those employers that *privately* report that they would definitely not hire the same worker in the future, *publicly* assign them 4 or more stars out of 5. The reverse essentially never happens—raters giving “good” private feedback and bad public feedback. The likely reason is that bad public feedback is costly in a way that bad private feedback is not. For example, employers may fear that workers who receive bad feedback will retaliate by complaining, bad-mouthing the rater, or withholding future cooperation. It is important to note that rated workers in our setting cannot retaliate by giving the employer bad feedback, because employer and worker feedback are revealed simultaneously (Bolton et al., 2013). The cost may even be altruistic, as employers may simply not want to harm the ratee’s future prospects, because public feedback is consequential. Together, these “reflected” costs make giving bad feedback costlier to the rater than giving good feedback.

The cost of giving bad feedback could explain why feedback scores are higher than they would be if more employers reported truthfully. However, it cannot by itself explain the dynamics of ever-higher scores we observe: inflation requires the cost of leaving a given bad score to increase over time. We hypothesize this is precisely what happens: as ratings become inflated, the same nominal feedback score (e.g., 2 “stars”) becomes costlier to the rated worker over time, and hence costlier for the rater to give, causing raters to further inflate their ratings. In other words, the cause of inflation is that what constitutes bad feedback—feedback that

⁴In addition to employing alternative measures of rater satisfaction to estimate the extent of reputation inflation, we also directly test alternative hypotheses that could explain the over time increase in feedback scores. In Appendix A, we provide tests that directly rule out several plausible explanations, including changes in employer, worker, and task composition, as well as trading partner selection. Furthermore, Appendix B verifies that the private feedback measure was not misinterpreted by the employers, and provides tests ruling out selection concerns in giving private feedback. Appendix C rules out selection and composition concerns in giving written feedback, and provides robustness tests for the predictive modeling approach.

causes worse market outcomes for raters receiving that score—is endogenous, depending on the current distribution of feedback scores, which in turn determines what inferences future buyers make from a score.

We formally illustrate how reputations can inflate through a simple model of a marketplace with a reputation system. We show that, in this model, there exists a unique and stable equilibrium in which reputations are universally inflated, because employers are more likely to report good feedback, regardless of actual worker performance, even when raters derive some benefit from telling the truth. For a simple matching process, and while holding the primitives of the model constant, reputation inflates over time following a pattern that is similar to the one we observe in our data. The degree of reputation inflation depends on how much cost the rated party that receives bad feedback can impose on the rater who gives it; this could explain why in less personal settings—such as consumers rating products on Amazon or restaurants on Yelp—ratings are more spread out. In contrast, inflation is likely more acute in highly “personal” settings, such as on peer-to-peer platforms (Einav et al., 2016; Filippas et al., forthcoming).

As a test for our model’s claim that reflected costs cause inflation, we report quasi-experimental evidence from a platform intervention that allows us to test the predictions of our model. After collecting private feedback for 10 months, the platform began publicly releasing *batched aggregates* of this private feedback score to future would-be employers, thereby making private feedback consequential to workers. This batching kept private feedback quasi-anonymous, because the worker would not know which particular employer gave which feedback (unless every rater in the batch gave the lowest or highest possible rating). Furthermore, as the platform chose to only display average private feedback scores for each worker, workers could not see previous private feedback scores assigned by employers.

The private feedback quasi-experiment created a positive shock to the cost of workers from receiving bad private feedback, allowing us to test our hypotheses that (1) the raters’ choice of what score to give is strategic—in the sense that employers consider the likely costs and benefits to what they report—and hence are more candid in private because there are no reflected costs, and (2) if employers have altruistic intents then inflation should occur, because the switch to public revelation made private feedback consequential to workers. To see why, it is important to note that this policy change would have different effects depending on the structure of these reflected costs. If employers only want to avoid worker retaliation, then we should observe no effect, because giving bad feedback to workers cannot “get back” to the employers—their identity remains hidden. However, if employers do not want to harm workers’ future prospects, then private feedback scores will start inflating. Of course, introducing a new, potentially more informative feedback measure may have caused improvements in fundamentals, and hence we still need to disentangle the effect of improvements in fundamentals from inflation on the observed changes, as we did with the whole history of feedback

scores.

We find that when the platform suddenly made private feedback scores public, private feedback scores began increasing immediately, but there was limited change in the sentiment of written feedback—the no-longer-private feedback became inflated, mirroring what we observed with public feedback. This result implicates the role of the reflected cost of bad feedback: when bad feedback became costlier to receive for workers, it also became costlier to give for employers, and there was less bad feedback. As bad feedback became scarce, what was mildly negative before became very negative, starting the inflation process described in our model.

Our paper makes several contributions. Our key contribution is documenting the extent of reputation inflation in a large online marketplace, and identifying its root cause. Our long-run, whole-system perspective is possible because we use data spanning over a decade of the operations of the marketplace. The reputation inflation problem is likely widespread, given that many online marketplaces share the same features as the one we study in depth, and nearly all have those features that we show lead to inflation. Our collection of longitudinal data from other platforms supports this view—in every marketplace for which we could obtain data, we observe average feedback scores increasing, even though none of these marketplaces allows “tit-for-tat” rating behavior (Bolton et al., 2013).

While our paper is not the first to explain how reputations can be biased, previous work has focused on cross-sectional data (Dellarocas and Wood, 2008; Tadelis and Zettelmeyer, 2015; Tadelis, 2016; Hu et al., 2017). We believe ours is the first paper to take a longitudinal approach and show how individually rational choices about what feedback to leave push the market towards a less informative equilibrium, as to establish that the ever-higher reputation scores are to a large extent due to reputation inflation.

Our analysis of the public revelation quasi-experiment highlights the key role of costs to raters, and that these costs are hard to diminish: making feedback quasi-anonymous failed to counteract inflation. Insofar that employers know bad feedback is consequential to workers, employers find it costly to assign bad feedback—even if the rated individual cannot retaliate. This suggests an inherent tension between ratings being consequential and ratings being informative. Whether reputation systems less prone to inflation can be designed remains an open research question.

The rest of the paper is organized as follows. Section 2 surveys related literature, introduces the empirical setting, and documents increasing feedback scores over time across five online marketplaces. Section 3 employs two alternative measures of rater satisfaction to show that the feedback score increase in our setting is to a large extent due to reputation inflation. To examine the causes of reputation inflation, Section 4 presents a simple theoretical framework, and Section 5 examines the quasi-experimental revelation of private feedback information. Section 6 concludes.

2 Empirical context

The setting for our study is a large online labor market. In online labor markets, employers hire workers to perform remote tasks, such as computer programming, graphic design, and data entry. Online labor markets differ in their scope and focus, but services provided by the platform are similar to those provided by other peer-to-peer markets, and include maintaining job listings, arbitrating disputes, certifying worker skills and, importantly, building and maintaining reputation systems (Horton, 2010; Filippas et al., forthcoming). Online markets offer a convenient setting for research due to the excellent measurement afforded in the online setting, and the ease of conducting field experiments (Horton et al., 2011).

2.1 Status quo reputation system

In our setting, when one party ends a contract both parties are prompted to give feedback.⁵ Employers are asked to give both written feedback, e.g., “Paul did excellent work—I’d work with him again” or “Ada is a great person to work for—her instructions were always very clear,” and numerical feedback. The numerical feedback is given on several weighted dimensions: “Skills” (20%), “Quality of Work” (20%), “Availability” (15%), “Adherence to Schedule” (15%), “Communication” (15%) and “Cooperation” (15%). On each dimension, the rater gives a score on a 1-5 scale.

The scores are aggregated according to the dimension weights. A worker’s reputation at a moment in time is the average of her scores on completed projects, weighted by the dollar value of each project. On the worker profile, a lifetime score is shown as well as a “last 6 months” score, which is more prominently displayed. Showing recent feedback is presumably the platform’s response to the opportunism that becomes possible once a employer or worker has obtained a high, hard-to-lower reputation (Aperjis and Johari, 2010; Liu, 2011). Despite the aggregation of individual scores into a reputation, the entire feedback “history” is available to interested parties for inspection. Workers can view the feedback given to previous workers rated by that employer, and the feedback received by an employer from that same worker.

The reputation system could be characterized as state-of-the-art, in the sense that direct tit-for-tat conditioning is not possible (Dellarocas, 2005; Bolton et al., 2013; Fradkin et al., 2019). Both the employer and the worker have an initial 14-day period in which to leave feedback. The platform does not reveal public feedback immediately, but rather uses a “double-blind” process. If both parties leave feedback during the 14-day feedback period, then the platform reveals both sets of feedback simultaneously. If only one party leaves feedback, then the platform reveals it at the end of the feedback period. Thus, neither party

⁵We use the present tense here to describe the reputation system before the introduction of private feedback. Although our focus is on employer-on-worker feedback, our claims carry through to the equally important case of worker-on-employer feedback (Benson et al., 2019).

learns its own rating before leaving a rating for the other party. Leaving feedback is strongly encouraged, but not compulsory. These encouragements seem effective, in that over the history of the platform, 81.8% of employers eligible to leave feedback have chosen to do so.

2.2 Feedback now and in the past

The distribution of employer-on-worker feedback scores in the market is highly right-skewed, but has not always been that way—scores have increased sharply over time. Most of the increase is explained by an increasing share of contracts receiving perfect 5-star feedback. These features of the data can be seen in the three panels of Figure 1.

Figure 1a depicts the histogram of public feedback scores from January 1, 2014 to May 11, 2016, for contracts worth more than \$10.⁶ Public feedback scores are between 1 and 5 stars, inclusive, and with increments of 0.25 stars. Each bar is labeled with the percentage of total observations falling in that bin, and the red dashed line shows the cumulative number of assignments with feedback less than or equal to the right limit of the bin it is above. More than 80% of the evaluations fall in the 4.75 to 5.00 star bin (1,339,071 observations). The average feedback pooled for the whole sample shown in Figure 1a is 4.77.

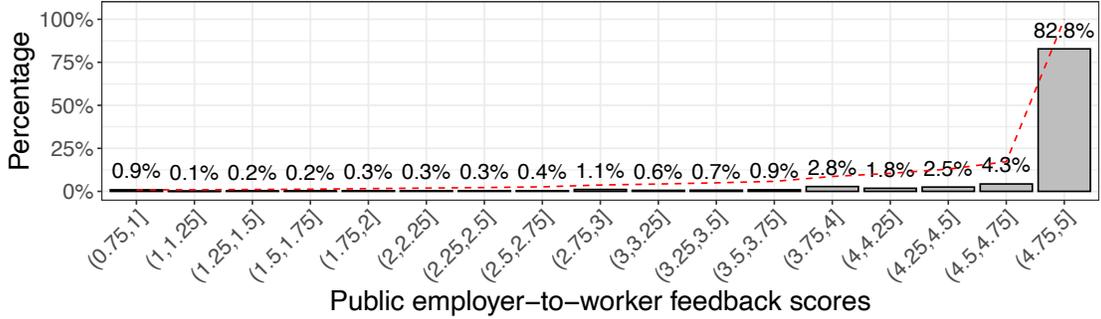
Scores have not always been highly right-skewed. Figure 1b shows the average monthly feedback over time, for contracts ending within each month. There is a clear increase in the feedback scores awarded on the platform: the feedback score average has increased from 3.74 in the beginning of 2007, to 4.85 in May 2016. The strongest period of increase was 2007, when average feedback scores increased by about 0.53 stars.

The increase in average feedback could be the outcome of raters giving less bad feedback, more good feedback, or some combination thereof. Figure 1c shows the fraction of contracts having a rating within different ranges, over time. In the early days of the platform rating assignments were reasonably dispersed, with completed contracts regularly receiving ratings in the (0, 3] range. Near the end of our data, completed contracts essentially never receive a rating in the (0, 3] range. Instead, there has been a dramatic increase in the fraction of contracts getting exactly 5 stars: 33% of contracts received a 5-star rating at the start of sample, compared to 85% at the end of the sample.

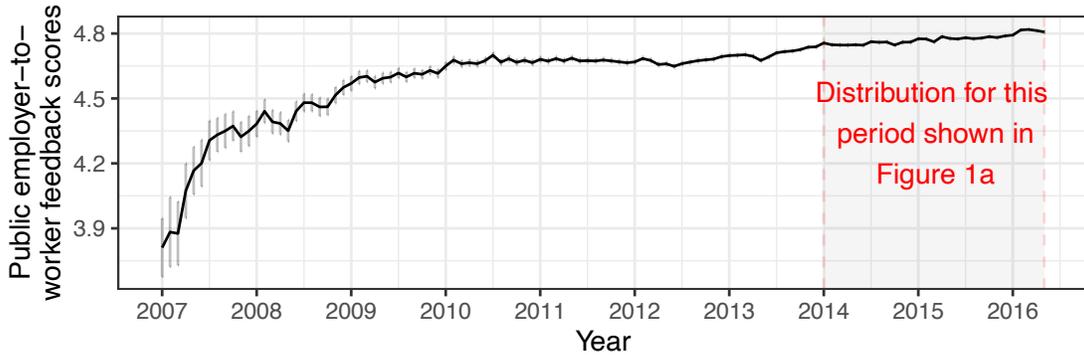
⁶We use this \$10 restriction throughout the paper to remove mistaken, trial, and erroneous transactions.

Figure 1: Employer-on-worker feedback characteristics in an online marketplace

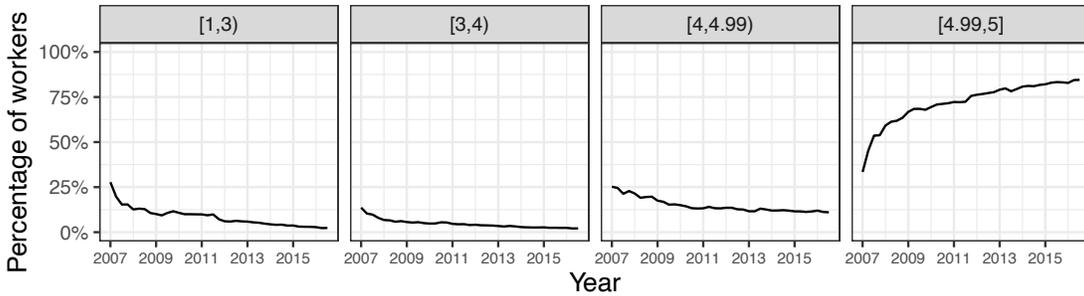
(a) Distribution of feedback scores for the period January 1, 2014 to May 11, 2016.



(b) Monthly average public feedback scores assigned on completed projects.



(c) Percentage of completed projects receiving different star ratings over time.



Notes: The top panel shows the histogram of public numerical ratings assigned by employers to workers, discretized by 0.25 star interval bins. The scale for feedback is 1 to 5 stars. The value of each bin is shown above it, and the red line depicts the empirical cumulative density function. The sample we use consists of all contracts worth more than 10\$ from January 1, 2014 to May 11, 2016, for which the employer provided feedback. See Section 2.2 for the description of the sample. The middle panel plots the average public feedback scores assigned by employers to workers on completed contracts by month. The average scores are computed for every month, and a 95% interval is depicted for every point estimate. The shaded area denotes the data that was used in Figure 1a. The bottom panel plots the fraction of public feedback scores assigned in a given month into four bins, [1, 3), [3, 4), [4, 4.99), and 5 stars, over time.

2.3 Evidence from other online marketplaces

Before exploring the causes for the pattern observed in Figure 1b, we turn to the question of whether the observed increase in feedback scores is specific to our setting or a general feature of reputation systems. Although there is substantial evidence of right-skewed distributions of ratings at a moment in time (Nosko and Tadelis, 2015; Athey et al., 2019), we are unaware of other research showing that this right-skewness arises over time rather than being present at launch. For this reason, we obtained data from four other online marketplaces, two of which contain city-level data.

The average feedback scores for the various marketplaces are shown in Figure 2. Panel (a) shows longitudinal data in a competing online labor market. Panel (b) plots longitudinal ratings data from four major cities in the United States and Europe in a large home-sharing platform. Home-sharing platforms are peer-to-peer marketplaces that facilitate short-term rentals for lodging (Filippas and Horton, 2018). Panel (c) plots numerical feedback data from an online marketplace that facilitates the short-term rental of a durable asset (Sundararajan, 2013; Filippas et al., forthcoming). Panel (d) plots longitudinal ratings data from six major cities in the United States in a large online marketplace for services.

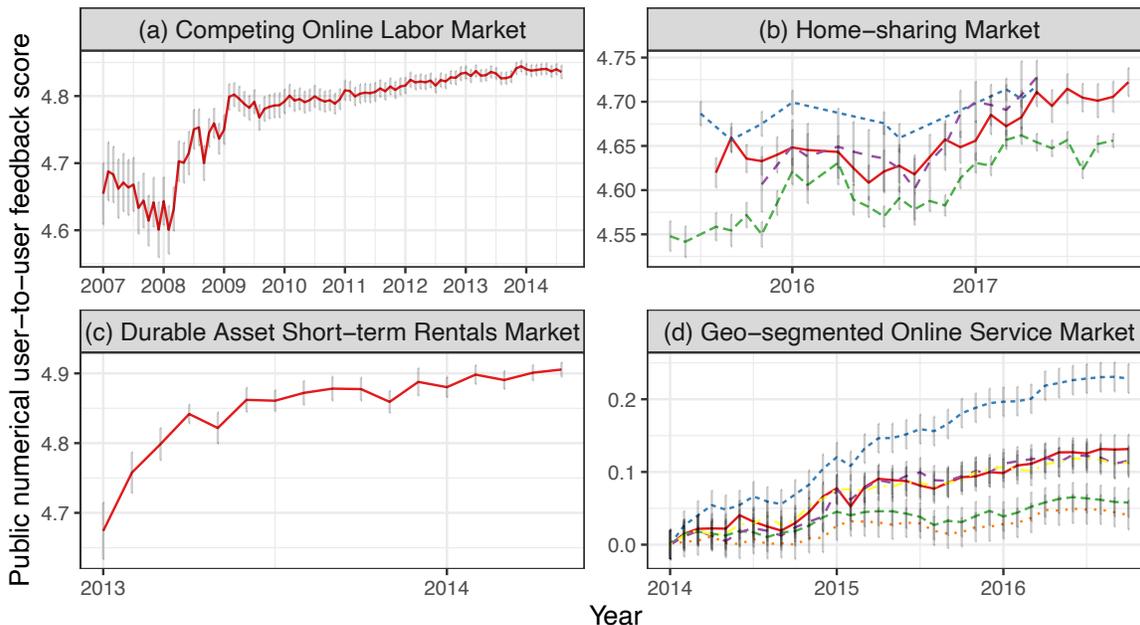
The goods and services that are transacted in these marketplaces differ dramatically. In panel (a) ratings are assigned by employers to workers, in panel (b) by guests (those who are renting properties) to hosts (those who are renting out properties), in panel (c) by users (renters of the asset) to users (providers of the asset) after the transaction has taken place, and in panel (d) by consumers of the service to providers of the service. Furthermore, these platforms greatly differ in the marketplace mechanisms they employ. For example, in the home-sharing marketplace renters choose the provider, but in the service marketplace the platform assigns a nearby available provider to the consumer upon request of the service.

Despite the differences in goods being transacted, transactions in these marketplaces are personal (peer-to-peer), and the same base reputation system is used. In particular, ratings are given after the transaction has taken place and are consequential for the rated party, and all platforms utilize mechanisms that prevent tit-for-tat rating behavior. We observe an increase in ratings over time that mirrors the pattern that we found in our setting. This provides us with evidence that increasing feedback scores are likely common in online marketplaces with similar reputation system characteristics, even when other marketplace characteristics vary, and irrespective of the transacted goods and services.

2.4 Related work on reputation systems

One particular focus of the online market design literature has been reputation systems. There is overwhelming evidence that reputation systems are economically significant, and confer substantial benefits for online markets by reducing moral hazard and adverse selec-

Figure 2: Longitudinal buyer-on-seller feedback scores for a collection of online marketplaces



Notes: This figure plots the average public feedback scores assigned in four online peer-to-peer marketplaces. In all markets, scores are assigned upon the completion of each transaction, and the scale for feedback is 1 to 5 stars. Scores are assigned by employers to workers in panel (a), by guests (users renting properties) to hosts (users renting out properties) in panel (b), by renters (those renting the durable asset) to providers (those renting out the durable asset) in panel (c), and by customers to providers of a service in panel (d). The lines in panels (b) and (d) correspond to different cities. In panel (d), average feedback scores for the time-series of each city are normalized so that the mean score is equal to zero during the first period of data collection. For each observation, average scores are computed for every time period, and a 95% interval is depicted for every point estimate.

tion (Resnick et al., 2000; Dellarocas, 2003; Jin and Leslie, 2003; Resnick et al., 2006; Cabral and Hortag̃su, 2010; Mayzlin et al., 2014; Ghose et al., 2014; Stanton and Thomas, 2015; Luca, 2016; Luca and Zervas, 2016; Benson et al., 2019). Our paper contributes to a recent stream within this literature that examines factors that may decrease the informativeness of reputation systems, and hence reduce their positive effects. The two main challenges in this line of research are (i) the difficulty of assessing the informational content of feedback scores, and (ii) identifying the reasons that lead to the informativeness decrease.

Assessing the informativeness of online reputation is closely related to the challenges faced by firms measuring customer satisfaction (Kamakura, 2010; Podsakoff et al., 2012; B̃uschken et al., 2013). To address these biases in a cross-sectional data, researchers have exploited sources of plausibly exogenous randomness, such as the availability of individual service employees (Huang and Sudhir, 2019). To bypass the same problem, we use a longitudinal

approach, measuring changes in feedback scores over time. Our approach is conceptually similar to estimating monetary inflation, and inflation in non-digital contexts (Sidrauski, 1967; Friedman, 1977; Diewert, 1998; Mishkin, 2000; Babcock, 2010; Berentsen et al., 2011). Klein et al. (2016) follow a similar approach to examine the market effects of a change in eBay’s reputation system.

To the best of our knowledge, our paper is the first to examine the informativeness of reputation systems over time. Several papers have examined the lack of information in feedback scores at a moment in time. The reasons that lead to less informative feedback scores have also received substantial attention (Avery et al., 1999; Dellarocas and Wood, 2008; Bolton et al., 2013; Nosko and Tadelis, 2015). Similar to the findings of our paper, Fradkin et al. (2019) also find a divergence between private and public feedback measures, which they attribute to reciprocal rating behavior due to the degree of social distance between the transacting parties. While reciprocity surely plays a role, our analysis of the private feedback revelation quasi-experiment supports that inflation occurs due to raters having an altruistic desire not to harm the rated party.

In addition to the difficulty of eliciting truthful feedback, recent work has also examined other factors that may reduce the informativeness of reputation systems. One threat to reputation systems is that feedback scores are prone to manipulation by malevolent users. Mayzlin et al. (2014) examine firms’ incentive to create fake reviews by exploiting the difference in reviews for a given hotel between Expedia, where only customers can post reviews, and TripAdvisor.com, where everyone can post reviews. In the same empirical context, Luca and Zervas (2016) analyze restaurant reviews that are identified by Yelp’s filtering algorithm as suspicious or fake, and find that such reviews have increased over time, and that restaurants generate fake reviews strategically to both harm competition and boost their own reputations. This kind of fraud is not possible in the reputation system of our setting, because (i) employers may only leave reviews only post-transaction, (ii) a worker’s reputation is the average of her scores on completed projects, weighted by the dollar value of each project, and (iii) a “last six months” score is shown along with the lifetime score.

Another reputation system feature that can bias feedback is allowing firms to respond to consumer reviews. Proserpio and Zervas (2017) examine how managerial responses to reviews affect hotels’ online reputations; they find that hotels start responding after a negative shock to their ratings, and that they subsequently receive fewer but longer negative reviews. Proserpio and Zervas argue that the mechanism driving this result is that unsatisfied consumers become less likely to leave short, indefensible reviews when hotels are likely to scrutinize them; once hotels start responding, they attract reviewers who are inherently more positive in their evaluations. In contrast, Chevalier et al. (2018) argue that negative reviews are more likely to be stimulated by managerial responses, because potential reviewers perceive negative reviews to be more impactful—a claim which the authors also corroborate

with experimental evidence. [Klein et al. \(2016\)](#) show that when eBay prevented sellers from negatively responding to bad buyer ratings, the reputation system became more informative, and buyer satisfaction increased.

In order to prevent feedback conditioning—tit-for-tat behavior and redress-seeking—feedback scores are simultaneously revealed and are immutable in state-of-the-art bilateral reputation systems of platforms such as Airbnb, Uber, as well as in our setting ([Bolton et al., 2013](#)). Despite this feature, the hardness of eliciting negative feedback persists: [Fradkin et al. \(2019\)](#) find that introducing simultaneously revealed Airbnb reviews reduced the percentage of 5-star ratings by only 1.2%.

Our paper also contributes to empirical studies of the long-run behavior of reputation systems, which has received little attention, perhaps due to the difficulty of obtaining longitudinal data. However, there has been some previous work leveraging longitudinal data, focused mainly on feedback scores in the context of online retail markets. [Li and Hitt \(2008\)](#) examine data from Amazon, and find evidence that consumers with higher valuations for a product are more likely to be early adopters, which can cause a decreasing trend in feedback scores. [Godes and Silva \(2012\)](#) find that the average Amazon review becomes more negative over time, which can be partially explained by a decrease in the diagnosticity of previous reviews. We highlight the need for a longitudinal view, documenting increases in feedback scores across five online peer-to-peer markets, and how the resulting top-censoring severely diminishes the informativeness of these scores. Our paper examines peer-to-peer markets, which we reason are distinctly different than retail markets.

3 Decomposing changes in reputation

The previous section documents a substantial increase in feedback scores over time across five online platforms. Two broad—and not mutually exclusive—sets of reasons may have led to substantially higher feedback scores: (1) rater satisfaction has increased, and (2) raters have lowered their rating standards.

Rater satisfaction can change over time, even if standards remain fixed. Reasons include platform improvements such as search and recommendation features resulting in better employer-worker matches, employers and workers becoming more experienced or exerting more effort, higher-quality cohorts of employers and workers joining the platform, low-quality employers and workers exiting the platform, employers incurring lower transaction and uncertainty costs by continuously transacting with a subset of desirable workers, and lower prices. Improvements in these marketplace “fundamentals” would result in higher average transaction quality, and hence higher feedback scores.

We cannot hope to account for all potential changes in fundamentals—and detect the increase in scores due to changes in standards. Instead, we can sidestep this issue by employ-

ing alternative measures of rater satisfaction. Let u_i denote the utility employer i obtains after some transaction. Employer utilities are unobservable, and potentially affected by both fundamentals-related and idiosyncratic reasons, which are also unobservable and may change over time. Upon receiving utility u_i , an employer leaves primary feedback

$$s_i = S(u_i) + \epsilon_s,$$

where $S(\cdot)$ is common among employers and monotonically increasing in the latent utility, and ϵ_s captures idiosyncratic differences in employers' rating behavior. For simplicity, we ignore discrete feedback scores, and instead assume that feedback scores are continuous scores with a noise component. Similarly, the same employer also leave alternative feedback

$$a_i = A(u_i) + \epsilon_a.$$

Assume that, at time t , utilities follow some latent distribution F . Let U denote a collection of observations, such that for all $u \in U$ we observe the measures of satisfaction $s_i(u)$ and $a_i(u)$, but we do not observe the actual utility u . We can use U to estimate a function \hat{s} that maps the alternative feedback measure a to the primary feedback measure s , i.e.,

$$\hat{s}(a_i(u)) = s_i(u) + \epsilon, \tag{1}$$

where ϵ is the estimation error. Essentially, \hat{s} estimates the conditional expectation $\mathbb{E}[s|a]$.

Assume now that at some latter period t' , employer satisfaction has shifted, i.e., the latent utility distribution has changed to F' , and that employers' rating standards have also shifted, i.e., the primary feedback measure has changed to $s'_i(u) = s_i(u) + \tau(u) + \epsilon_s$, where $\tau(u) > 0$ is the inflationary component. Although we no longer observe s and instead only observe s' , the mapping \hat{s} allows us to recover what the average value of the feedback measure s would have been if rater standards had not shifted. We get

$$\begin{aligned} \Delta_s &= \int \underbrace{s'_i(u) - \hat{s}(a_i(u))}_{\text{actual - predicted score}} dF'(u) \\ &= \int \tau(u) - \epsilon dF'(u) \\ &= \underbrace{\mathbb{E}_{F'}[s'] - \mathbb{E}_{F'}[s]}_{\text{reputation inflation}} - \mathbb{E}_{F'}[\epsilon]. \end{aligned} \tag{2}$$

Insofar as $\mathbb{E}_{F'}[\epsilon] = 0$, we can estimate the extent of reputation inflation under the utility distribution F' . Crucially, if the alternative measure a is also subject to inflation or if

$\mathbb{E}_{F'}[\epsilon] > 0$, then the estimate Δ_s provides a *lower bound* for the inflation term.

Using an alternative feedback measure to estimate reputation inflation circumvents the problem of estimating latent utilities from observational data, and hence accounts for improvements in marketplace fundamentals.⁷ Instead, our approach presents us with different—albeit more tractable—problems. In particular, our estimation method requires us to verify two conditions: (i) that a one-to-one mapping exists between the latent utility and the alternative feedback measure, and (ii) that the term $\mathbb{E}_{F'}[\epsilon]$ is non-negative.

The first condition guarantees that we can use the law of total expectation in Equation 2 to compute $\mathbb{E}_{F'}[s] = \int \mathbb{E}[s|a = a(u)]dF'(u)$. To see why, assume that a one-to-one mapping does not exist, and that there are some u_1, u_2 such that $u_1 > u_2$ and $a(u_1) = a(u_2)$, but $s(u_1) \neq s(u_2)$. Because utilities are latent and we only observe $a(u_1)$ and $a(u_2)$, our procedure may then fail to recover $\mathbb{E}_{F'}[s]$ correctly. We assume that this is not an issue in our context, as measures of employer satisfaction are increasing in the latent utility, and hence the function \hat{s} is injective. For the second condition, the estimation error $\mathbb{E}_{F'}[\epsilon]$ term can be non-zero due to (i) differential selection of latent utilities of which \hat{s} over- or under-estimates in expectation, and (ii) because of systematic changes in how raters translate their utilities to the primary and alternative feedback measures. The former condition is directly testable, while the latter needs to be reasoned through theoretically.

In what follows, we employ “private” and written feedback as two alternative measures of employer satisfaction. For both measures, we will reason why they are not affected by systematic biases, and hence why they are subject to less inflationary pressure. In addition to our alternative measures approach, we provide direct tests in Appendix A that show that the most plausible selection and composition stories cannot explain the observed increase in feedback scores.

3.1 Private feedback as an alternative measure

Our first alternative measure of rater satisfaction comes from a platform experiment that elicited an additional private feedback measure. At the completion of a contract, employers

⁷Our decomposition task is conceptually similar to estimating monetary inflation in the presence of quality changes (Sidrauski, 1967; Friedman, 1977; Mishkin, 2000; Berentsen et al., 2011). Although quality changes are acknowledged, they are typically sidestepped by a “basket-of-goods” approach (Diewert, 1998). The implicit assumption underlying such methods is that consumers derive the same satisfaction from some “basic” goods and services, irrespective of the time period—a consumer derives equal utility from a loaf of bread in 2000, as she will in 2020 (approximately, and, of course, not from the same loaf of bread). Issues with such measures of monetary inflation mostly arise when aggregate consumer utility from basic goods and services changes and is hard to measure, such as for goods including phones, computers, and even cars. In the context of online markets, however, there is no basic or standard transaction, i.e. a transaction with an immutable associated rater satisfaction. Our approach builds on Klein et al. (2016), who examine the differences between two measures of user satisfaction on eBay. Other approaches—which we do not take in this paper—would be to debias consumer satisfaction estimates directly (Huang and Sudhir, 2019), or to estimate structural models of the value of reputation across different time periods (Yoganarasimhan, 2013).

were asked for private feedback, in addition to public feedback. Critically, the platform let the employers know that private feedback would not be shared with other workers or employers, and would only be collected by the platform for internal evaluation purposes. As it was less costly for workers, this private feedback offers an alternative measure that is potentially less subject to inflation.⁸ Employers were initially asked the private feedback question, “Would you hire this freelancer[worker] again, if you had a similar project?” Starting on June 2014, employers were instead asked to rate workers on a numerical scale of 0 to 10, answering the question “How likely are you to recommend this freelancer to a friend or colleague?”

Figure 3 shows the distribution of public feedback, conditioned on the private feedback. The percentage of employers giving that feedback score is shown in parenthesis in each panel. Although the most common response was “Definitely Yes,” about 15% of the employers gave unambiguously bad private feedback (“Definitely Not” and “Probably Not”). In contrast, during the same period less than 4% of the employers gave a numerical score of 3 stars or less. Given this gap, we might suspect that some employers expressing a negative private sentiment are less candid in public.

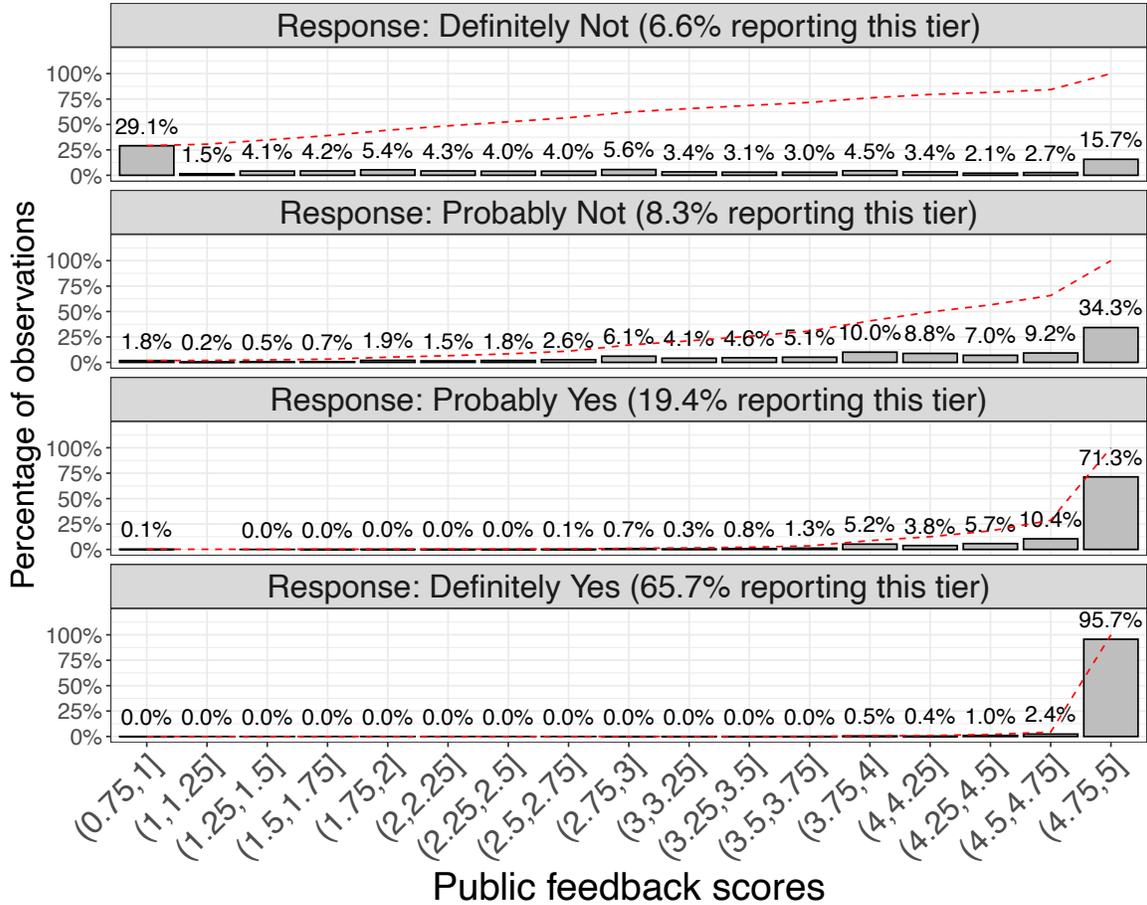
Employers who leave more negative private feedback do assign lower public feedback scores: among those employers that selected the “Definitely No” answer to the private feedback question, 29.1% assigned a 1-star rating publicly. Surprisingly, however, the second most common choice for these employers at 15.7% was in the 4.75 to 5.00 bin, and 28.4% publicly assigned more than 4 stars. In short, many privately dissatisfied employers publicly claimed to be satisfied. We can see that the reverse—privately satisfied employers giving bad public feedback—essentially never happens. Employers who selected “Definitely yes” left very positive public feedback, with more than 95% of these observations falling into the highest bin.

Figure 4a reports the average monthly feedback over time, for the numerical public and private feedback during the period that both were collected. To make the two scores comparable, we normalize them by the first observed mean. In the language of Section 3, we use $\hat{s}(a_t) = (a_t - a_0)/a_0$, which has the advantage of simplicity. Public feedback scores exhibit a small increase during the period of interest (as we saw in Figure 1b). However, for the same period of time, private feedback scores exhibit a strong decreasing trend; overall, the divergence in the two scores at the end of the 9-month period is 3.5 percentage points.

It is critical to stress that the average feedback scores shown in Figure 4a are being assigned by the same employers on the same contracts. The decreasing private feedback scores hence suggest a decline in rater satisfaction, and yet public feedback increased. In

⁸Even if employers believed that the platform would use the private feedback information—for example, to suggest areas of improvement to workers—private feedback was not publicly displayed, and hence was less consequential for worker’s future outcomes than public feedback. As such, private feedback was subject to less inflationary pressure than public feedback was—we elaborate on this argument in Section 4.

Figure 3: Distribution of public employer-on-worker feedback, by employers’ response to the private feedback question: “Would you hire this freelancer [worker] again, if you had a similar project?”

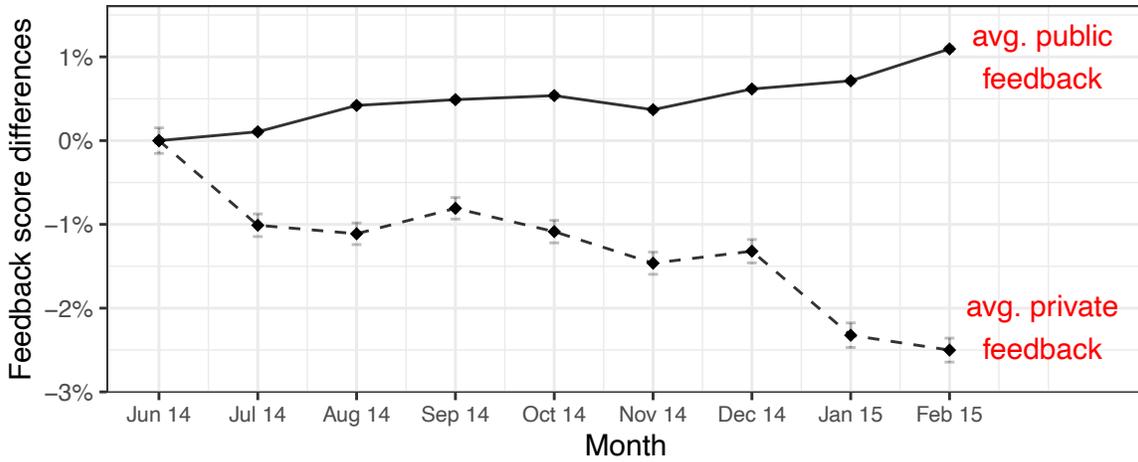


Notes: This figure plots the distribution of public feedback scores, computed separately for every set of users that gave the same answer to the private feedback question. The red dotted line plots the cumulative distribution function.

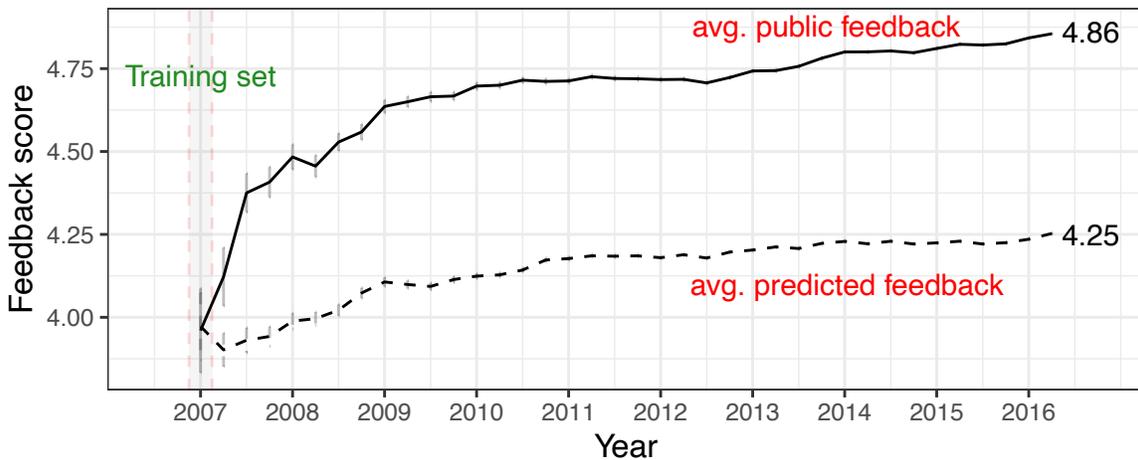
Appendix B, we provide more information on how private feedback was elicited, as well as a series of robustness checks that rule out other assumptions that could rationalize the divergent trends—e.g., that private standards are getting harsher although transaction quality is increasing, or that there is a substitution effect from public to private feedback scores. As such, it is hard to rationalize some change in fundamentals alone that could generate this pattern. What seems more probable is that public feedback scores are subject to inflation, whereas the private scores are not because of their private nature.

Figure 4: Using alternative measures of rater satisfaction to quantify reputation inflation

(a) Numerical public feedback score and private feedback score.



(b) Numerical public feedback score and predicted score from textual feedback



Notes: This figure presents evidence of reputation inflation in public feedback scores by employing alternative measures of rater satisfaction. The top panel shows the evolution of the average public feedback scores (solid line) versus the average private feedback scores (dashed line) assigned by employers to workers, for the same contracts. The average scores are computed for every month, and are normalized by the value of their respective first observation (June 2014). A 95% confidence interval is shown for each mean. The bottom panel shows the evolution of average public feedback scores (solid line) versus the average predicted score of textual feedback (dashed line) assigned by employers to workers. A 95% interval is depicted for every point estimate. The shaded area indicates the quarters from which training data was obtained for the corresponding predictive model.

3.2 Written feedback as an alternative measure

Our second alternative measure of rater satisfaction comes from written feedback. In contrast to private feedback, we have written feedback over the entire platform history. Similar to the private feedback, there are several reasons that the costs to the rater for giving negative written feedback are lower than for numerical feedback. First, it is harder for workers to complain about textual tone than it is to complain about a non-perfect star rating. Furthermore, the platform does not aggregate written feedback or put it on a scale, and the written feedback history is not presented in the worker’s profile page which is typically accessed by employers during the initial worker screening phase—only average numerical feedback scores are presented, and written feedback is harder to access. As such, written feedback cannot be used for cross-worker comparisons by future employers as easily as numerical feedback does; these comparisons are precisely what makes feedback consequential for workers.

To make the two kinds of feedback comparable, we fit a predictive model, $\hat{s}(\cdot)$ that predicts numerical feedback scores from the feedback text. The predictive model is fit on a narrow time window, using the written feedback corpus as the training set and the associated numerical scores as the set of labels. Each written feedback left by an employer post-transaction is one instance in our data. To learn the predictive model, we use a standard natural language processing pipeline. For the preprocessing step, the text of each employer-on-worker review is stripped of accents and special characters, is lowercased, and stopwords are removed. A matrix of token counts (up to 3-grams) is created, and is weighed using the TFIDF method. To find the best-performing algorithm, we conduct an extensive grid search, evaluating each configuration of hyper-parameters using a 5-fold cross validation in terms of average squared error. We then use the fitted model to estimate out-of-sample feedback scores of the written feedback for the entire sample.

The average quarterly feedback scores over time, for both the numerical public feedback, and the feedback predicted from the written feedback, are plotted in Figure 4b. As expected, the two scores match up during the training period. Going forward, both scores increase, but the predicted feedback score increases at a much slower rate. On average, numerical feedback goes from 3.96 in the beginning of 2006 to 4.86 stars at the beginning of 2016. In contrast, the average score predicted from the written feedback only goes to 4.25 stars.

The divergence between written sentiment and numerical feedback implies that a substantial amount of the increase in numerical feedback scores is due to lower rater standards. Our approach also allows us to quantify the degree of inflation: the point estimate is that 67.7% of the increase in feedback scores is due to inflation.

Importantly, written feedback can certainly become inflated, with work that would have elicited a “good” now garnering a “great.”⁹ We have some evidence that written feedback

⁹One written feedback in our data reads: “This is the most impressive piece of coding in the history of

inflates, in that the private feedback scores were declining while the sentiment of written feedback was increasing (see Figure 4). Regardless, to the extent that written feedback is also subject to inflation, our approach will *underestimate* the magnitude of the inflation in scores, and hence our estimates can be interpreted as lower bounds.

Our approach requires the assumption that there is no selection with respect to bias in the model or the rater. Although this assumption is not directly testable, in Appendix C we report a number of tests looking for evidence of selection bias with respect to the written measure, finding no evidence against our assumption. Furthermore, in Appendix C we also verify that different training periods and/or predictive algorithms yield similar results.

A potential threat to our approach is that the lexical composition of reviews could presumably change over time; in the language of our model, $a(u_i)$ has shifted to $a(u_i) + h(u_i)$. While we have no evidence that supports this hypothesis, in what follows we take an alternative approach: as a more direct measure of inflation from written feedback, we can examine whether the same sentences found in written feedback correspond to different feedback scores at different points in time. The advantage of this approach is that it does not depend on the choice of predictive model.

We select written feedback from 2008 and 2015, and find all lexically identical sentences generated in these periods. We then compare average feedback by sentence, across the two periods. To illustrate this approach, Figure 5 shows the average numerical feedback scores for a set of commonly used short sentences, by period. We select sentences spanning both good and bad feedback, and which most frequently occurred in the corresponding written feedback in our data. Across terms, we see that the numerical feedback scores associated with identical sentences have increased considerably over time, and that this increase has affected both positive and negative sentences.

4 A model of reputation inflation

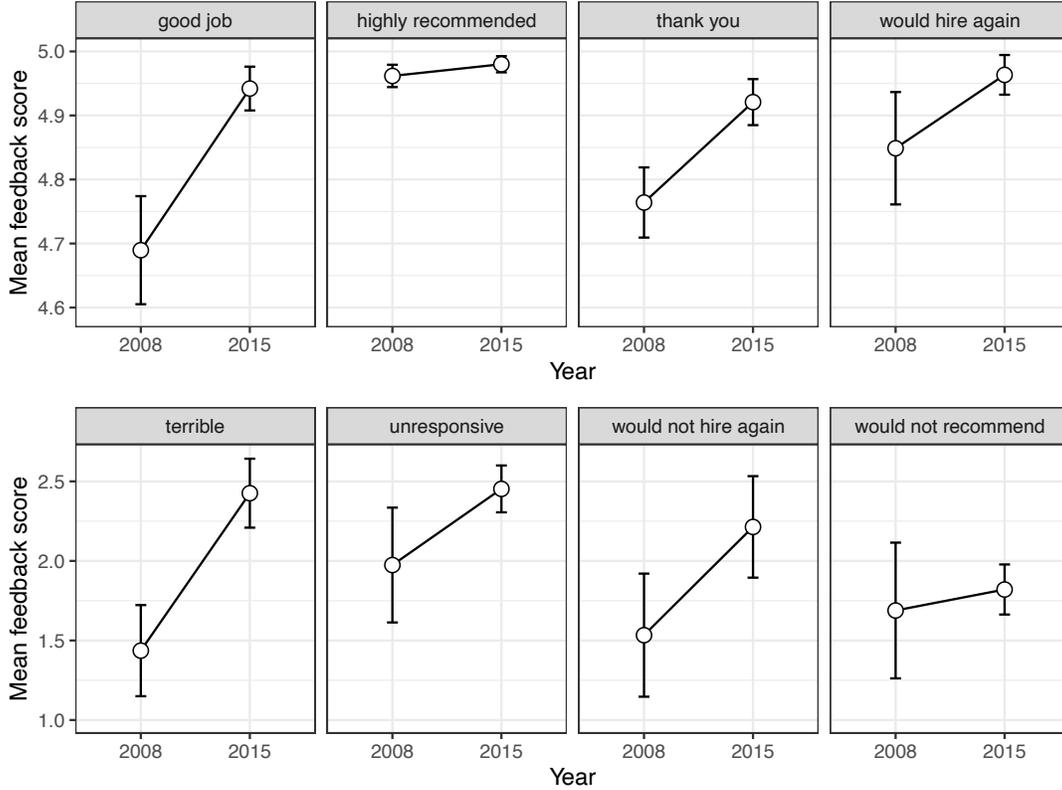
We next provide a framework to help explain the implications of our empirical findings, and to examine *why* reputation inflation occurs. Toward that end, we develop a model of reputation inflation in the reputation system of a competitive market. Although our motivating example is a labor market, our framework can be applied to the more general case of markets where participants assign public post-transaction feedback.

The focus of our model is the employers’ decision to either assign bad feedback candidly following a bad experience, or to lie and assign good feedback. This assumption is motivated by the strategic misreporting we observe in our data—reviewers choose to leave good public feedback despite an unsatisfactory experience, at least as they stated in private.¹⁰

software development!”

¹⁰Note that employers essentially never strategically misreport a good private experience in our data, that

Figure 5: Difference over time in the feedback scores associated with identical sentences.



Notes: This figure shows the average numerical feedback associated with identical sentences found in the text of employer-on-worker written feedback, in 2008 and 2015. The sentences plotted are the four most common sentences associated with high feedback scores, and the four most common sentences associated with low feedback scores. A 95% confidence interval is shown for each mean.

In our model, employers have an incentive to truthfully assign bad feedback after a bad experience, captured by a positive benefit associated with truth-telling. This benefit includes idiosyncratic reasons to report truthfully as well as platform-specific benefits, such as awards by other users for being an accurate reviewer.¹¹ At the same time, employers also incur a cost when they assign bad feedback, which is increasing in the workers’ cost from receiving this bad feedback. This “reflected” disutility includes the employer’s aversion to harming the worker’s future prospects, the cost of the worker complaining or withholding future cooperation, and even the cost from other workers being unwilling to work for the employer in the future if the employer has a reputation as a strict rater.

is, employers never report good private feedback but bad public feedback. As such, we restrict employers’ choices in our model to whether a bad private experience should be publicly reported or not. For a comic strip on both types of misreporting, see <http://xkcd.com/958>.

¹¹Abeler et al. (2019) find strong evidence about individuals’ preferences for truth-telling, both in 90 previous studies and in their experiments. Surprisingly, the propensity for truth-telling persists even in one-shot games.

4.1 Setup

Consider an online labor market composed of workers and employers. Workers are matched at random with employers, after which workers produce output $y \in \{0, 1\}$. The worker produces output $y = 1$ with probability $\Pr(y = 1|q) = q$, from which the employer obtains utility equal to 1, by selling the output on some product market. The employer obtains zero utility in the case that output $y = 0$ is produced.

Workers are characterized by their quality $q \in \{q_L, q_H\}$, with $q_L < q_H$. Employers know the fraction of high quality workers in the marketplace, which we denote by θ . After the employer observes the worker's realized output y , she generates a signal to the marketplace in the form of feedback $s \in \{0, 1\}$, where $s = 1$ denotes good feedback, and $s = 0$ bad feedback. In the next round, employers observe the most recent feedback assigned to the worker, and form Bayesian beliefs about the worker's quality. We assume that both sides are price-takers, and hence workers are paid their expected marginal product, which is

$$w_s = \Pr(q = q_H|s)q_H + (1 - \Pr(q = q_H|s))q_L.$$

The worker's cost of bad feedback, realized in the next round, is the difference in compensation between receiving good feedback, $w_{s=1}$, and bad feedback, $w_{s=0}$, that is

$$\Delta w = w_{s=1} - w_{s=0}.$$

Whenever the employer tells the truth, that is when $s = y$, she obtains a benefit $b > 0$. If the worker's output is good ($y = 1$), then the employer has no incentive to lie and always assigns good feedback ($s = 1$) to the worker. However, in the case that the worker produces bad output ($y = 0$) and the employer truthfully reports $s = 0$, the worker incurs a cost Δw , which is the wage penalty in the next round.

We assume that some fraction of the cost Δw is reflected back on the employer. Employers differ in how much of this cost is reflected: let c_i be the employer-specific fraction of the worker's cost that is reflected back on the rating employer. The employer thus incurs a cost of $c_i \Delta w$, where c_i is drawn from some distribution $F : [\underline{c}, \bar{c}] \rightarrow [0, 1]$, with $\underline{c} \geq 0$.

In light of these reflected costs, some employers might give positive feedback even if the worker's output was bad, thereby avoiding the cost of giving bad feedback. This decision will depend on c_i , and so employer i will not report truthful feedback if

$$b \leq c_i \Delta w. \tag{3}$$

Let p denote the fraction of employers that generate truthful feedback in the most recent round, and assume that p is common knowledge. When considering a particular worker that

received bad feedback in the previous round, i.e., $s = 0$, the Bayesian employer infers that

$$\begin{aligned}\Pr(q = q_H | s = 0; p) &= \frac{\Pr(s = 0 | q = q_H; p) \Pr(q = q_H)}{\Pr(s = 0; p)} \\ &= \frac{(1 - q_H)\theta}{(1 - q_H)\theta + (1 - q_L)(1 - \theta)}.\end{aligned}$$

Note that the p term divides out as $s = 0$ always implies truthful reporting. In contrast, if the worker received good feedback, i.e., $s = 1$, the Bayesian employer infers that

$$\begin{aligned}\Pr(q = q_H | s = 1; p) &= \frac{\Pr(s = 1 | q = q_H; p) \Pr(q = q_H)}{\Pr(s = 1; p)} \\ &= \frac{(q_H + (1 - q_H)(1 - p))\theta}{(q_H + (1 - q_H)(1 - p))\theta + (q_L + (1 - q_L)(1 - p))(1 - \theta)}.\end{aligned}$$

The cost of bad feedback to a worker is then

$$\Delta w(p) = w_{s=1;p} - w_{s=0;p} = \frac{\theta(1 - \theta)(q_H - q_L)^2}{k - pk^2}, \quad (4)$$

where $k = \theta(1 - q_H) + (1 - \theta)(1 - q_L)$, which is the probability that a randomly chosen worker will produce bad output.

We see from Equation 4 that $\Delta w(p) > 0$ for all p , implying that as long as $c_i > 0$, employers incur a cost for giving bad feedback, which they must compare to their truth-telling benefit b . Furthermore, when p is large—when most of the employers report truthfully—feedback is a more accurate measure of quality, and hence the value of positive feedback increases, along with the wage penalty $\Delta w(p)$. In contrast, when the majority of firms lie, the signal from good feedback is less informative, and the wage penalty narrows, as many workers receiving good feedback actually produced “bad” output.¹²

The relationship between the wage penalty and the truth-telling fraction p makes which feedback is good and bad endogenous in our model—the characterization depends on p , which in turn depends on the choices of all other employers, who are reacting to that wage penalty. We now consider what an equilibrium of this market would be. Let p_E denote the fraction of firms that truthfully assign negative feedback when the market equilibrium has been attained. The equilibrium truth-telling fraction is found by solving the equation

$$p_E = F\left(\frac{b}{\Delta w(p_E)}\right). \quad (5)$$

An equilibrium exists for any continuous distribution function, and the equilibrium is unique

¹²In Appendix D we examine how alternative assumptions on output costs and on the employers’ rating behavior affect the relationship between the wage penalty Δw and the truth-telling fraction p .

for distributions with increasing hazard rates. The two extreme cases where

$$p_E = \begin{cases} 1, & \text{if } b \geq \bar{c}\Delta w(1) \\ 0, & \text{if } b \leq \underline{c}\Delta w(0) \end{cases},$$

correspond to the all-truthful and an all-lying equilibria. If the benefit to assigning truthful feedback is higher than the cost for every employer, then no employer has incentive to lie ($p_E = 1$), while if the costs are too high, all employers lie ($p_E = 0$).¹³ To the extent that we think of employers as both strategic and narrowly self-interested, the all-lying equilibrium is the likely equilibrium, as the benefit b is likely small or sometimes even zero, while the employer-specific costs c_i could be substantial.

4.2 Discussion of model results

Our model gives reflected costs a large role. These reflected costs can explain the divergent public and private feedback trends observed in Section 3.1: private feedback scores are not observable by other employers when they are making hiring decisions, and hence receiving bad private feedback is less costly for workers. As a result, employers are more truthful in private, and less truthful in public. In Section 5, we test the reflected cost assumption empirically, and further examine its underlying mechanisms.

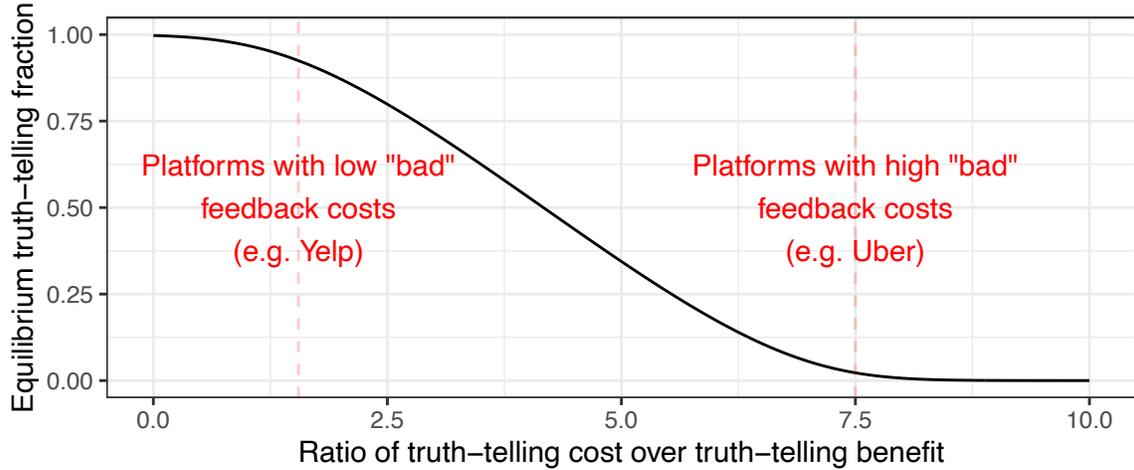
One prediction of our model is that reputation inflation will be acute when workers' costs from receiving bad feedback are high, and hence employers' reflected costs are high as well, thereby inducing employers to be less truthful in equilibrium. Specifically, our model predicts that reputation inflation will be severe in peer-to-peer platforms, such as online labor and sharing economy markets, where wage penalties for workers and employers' reflected costs are high. Some reasons for the higher worker cost from receiving negative feedback include that feedback scores are often the sole signal of quality, workers are typically highly substitutable and have few transactions, and hence each rating is more consequential. As transactions are more "personal," the reflected costs for employers are also likely higher. In contrast, when individuals assign feedback to products (e.g., movie reviews) there is likely no reflected cost, and inflation will be less acute. Similarly, institutional ratings—such as BBB and health inspection scores—are also less likely to suffer from inflation.

To provide a graphical depiction of this intuition, we plot in Figure 6 the truth-telling equilibrium fraction for different truth-telling cost to truth-telling benefit ratios. When the cost-to-benefit ratio is low, most employers report their feedback truthfully in equilibrium. This is the case for platforms such as Yelp or movie rating websites, where the raters are giving feedback to businesses, and transactions are less personal. Furthermore, reviewers in

¹³In the case where all employers have the same cost truth-telling cost, p_E can be interpreted as the probability of truthfully generating public negative feedback in the resulting mixed strategy equilibrium.

these platforms likely view themselves as performing a service for fellow consumers, and being known for good, honest reviews is at least part of the incentive people have for participating. In the language of our model, these sites have a higher b . As the cost-to-benefit ratio increases, the equilibrium truth-telling percentage approaches zero, and the average feedback scores are inflated: this is the case for platforms such as Uber or Airbnb, where the reflected costs are high.

Figure 6: Truth-telling equilibrium fraction as a function of the ratio of the mean cost over the benefit of generating truthful feedback.



Notes: This figure plots the equilibrium truth-telling fraction p_E as a fraction of the ratio of the average truth-telling cost μ_C over the truth-telling benefit b . The parameters used in computing the equilibrium truth-telling fractions are $q_H = 0.8$, $q_L = 0.2$, $\theta = 0.5$, $b = 1$, and $F \sim N(\mu_C, 1)$. To vary truth-telling costs, we vary the mean μ_C of the distribution of the reflected cost coefficients, keeping everything else constant. Alternative distribution and parameter choices yield qualitatively similar results.

We examine extensions to the base model in Appendix D. We first study the equilibrium convergence process under a simple matching scheme, where employers and workers match and transact randomly in every round. We find that reputation initially inflates fast, but the rate of inflation flattens out as the equilibrium truth-telling fraction is approached. This finding mirrors the pattern we observed empirically in all marketplaces we have data spanning their entire operations. Second, we show that alternative assumptions on how employer assign ratings can change how worker types “pool” into the different feedback score levels, and examine how this departure from the base model changes the relationship between the wage penalty Δw and the truth-telling fraction p .

5 Effects of making feedback consequential

The framework of Section 4 proposes a process by which reputation inflates. A key feature of the framework is that an employer misreports following a bad experience if $b \leq c_i \Delta w$ (see Equation 3). If the cost of bad feedback to the workers is zero, then employers should be truthful for any positive value of b , and thus would generate more bad feedback. If the cost of bad feedback to the workers changes, then the fraction of truthful employers should also change. Furthermore, as the cost of bad feedback is endogenous, our model also predicts a convergence to new equilibrium following a change in costs. As such, in the event that the costs of assigning bad feedback change, we do not expect to see a “jump” to the new equilibrium, but rather a gradual convergence to some new equilibrium.

Recall from Section 3.1 that employers were more candid about bad performance in private than they were in public. Our interpretation of public versus private feedback is that for bad public feedback, the cost to the worker, Δw , was higher than for the cost of bad private feedback. As a result, private feedback was more candid, i.e., more employers were more likely to report $s = 0$ when $y = 0$, as the employers’ costs were increasing in the workers’ cost of negative feedback.

We now consider what happened when the platform made a change that raised the cost Δw of assigning bad private feedback. The change was the platform’s announcement in March 2015 that the private feedback ratings would be used to compute a new aggregate feedback score for workers. The aggregate score on a worker’s profile was only updated after the worker received five new feedback scores, to prevent workers from identifying which employer gave them which feedback. This score would be shown on the profile of each worker and therefore be publicly available, but anonymous in the sense that workers could not associate individual scores with employers.

To the extent that employers used this new score in their hiring decisions, the workers’ cost of bad private feedback increased. In the logic of our model, the platform’s hope was that by not allowing workers to know which employer gave feedback, the distribution of c would remain unaffected and close to zero, even though Δw increased. However, if many employers simply do not want to hurt the worker or fear some other kind of generalized ex post retaliation, then even the batched release of private feedback scores would keep the weight of the distribution of c above 0, which should cause the private feedback measure to inflate from the all-truthful equilibrium to some new equilibrium.

Of course, simply observing that the time series of numerical private feedback rises after the platform change does not prove inflation. The new feedback system was intended to improve matches, and so the same concern from our earlier analysis applies—namely that any increase in the private feedback score following revelation reflects changes in fundamentals. For example, if employers could now form better matches because of their access to the

private feedback score measure, then we would expect higher future private feedback scores. As before, we address that concern by using written comments to construct an alternative measure of employer utility.

Figure 7 plots the monthly average private feedback and the monthly average *predicted* private feedback for transactions that received both types of feedback, using the same modeling method we used earlier in Section 3.2. In each panel, the monthly averages are shown by type, as well as the fitted values under different regression specifications. The predicted private feedback is the prediction of a model trained during a period before the revelation, which is indicated with a shaded region. The day the platform switched to batched public revelation is indicated with vertical dashed line. The figure shows that prior to public revelation, the actual and predicted private feedback are quite similar, but that after public revelation, the actual numerical rating increases while the predicted rating does not.

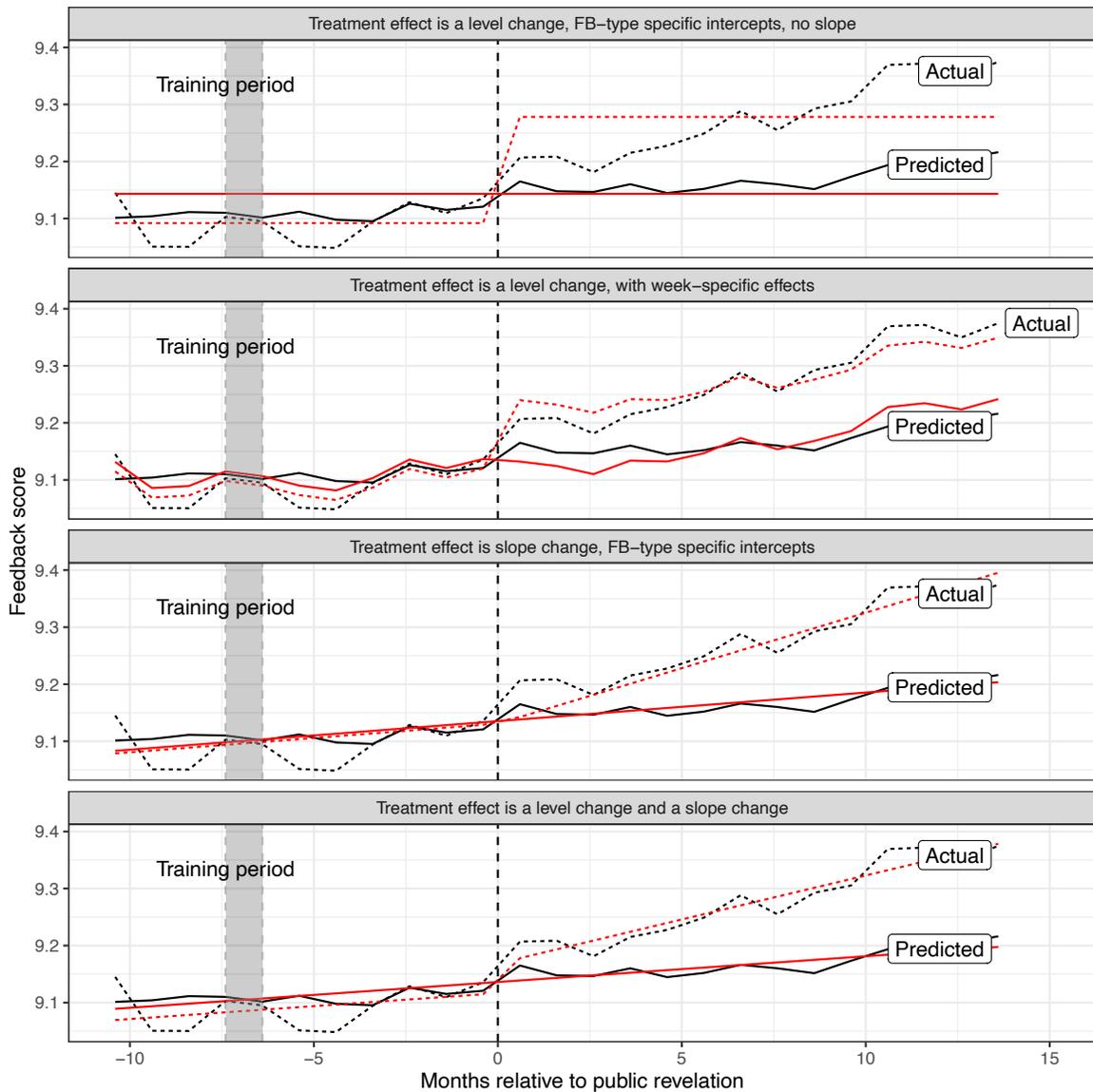
To quantify the effect of revelation, we switch to a regression framework. However, as we have some choice over the regression specification, the different panels of Figure 7 show various alternatives. In the top panel, we report the simplest specification, which is for the treatment to simply have a level effect and to allow the two feedback-types to differ by a fixed amount before the change. We can see that this specification clearly fails to capture the underlying time trend in both series, and especially for the numerical feedback in the post period. In the next panel down, the specification maintains the assumption of a level feedback, but includes a week-specific effect. This specification better captures the underlying trend in both measures that caused the previous specification to perform poorly, but it still performs inconsistently in the post-period, over-estimating the actual feedback early in the period, and then under-estimating it later, and vice versa for the predicted feedback. This is consistent with the simple level-change specification not capturing some of the dynamics of the effects of the treatment e.g., a change in slopes.

In the third panel from the top, we give both types of feedback a common linear time trend, but then allow that trend to change in the post-period for the actual feedback. With a common slope, the fit in the pre-period is much better than when we forced the two types to only differ by a level (in the top panel). However, we can see that earlier in the pre-period, only allowing a linear change in slopes under-predicts the actual feedback score, suggesting some immediate effect and not just a change in slopes.

In the bottom panel, the specification allows for both a level treatment effect and a change in slopes. This specification seems to work the best, with the predicted series closely matching the realized value. We will make use of this insight when we switch to estimating the effects of public revelation at the level of the individual contract rather than at the level of monthly averages. This has the advantage of allowing us to directly control for employer-specific effects and thus directly control for some of the potential sources of bias.

As our interest is in the divergence between the public and private feedback scores, we

Figure 7: Monthly average private feedback scores and average predicted private feedback scores



Notes: This figure shows the average monthly private feedback (on a 1 - 10 point scale) given by employers to workers, both actual and predicted. Predicted scores are derived from the employer's written textual public feedback, with the predictive model fit using data from the shaded region. The vertical line indicates the point in time in which employer private feedback scores were aggregated and added to worker profiles. These aggregate scores were changed after the worker received five new feedback scores, to prevent workers from identifying which employer gave them which feedback. Prior to this point, scores were only collected by the platform and not used publicly in any way. The red lines in the lower panels correspond to predictions from various difference-in-differences model specifications.

switch our outcome to Δs_i , which is the numerical private feedback rating minus the predicted private rating based on the sentiment of the written text. By taking this difference, we eliminate the need (or the possibility) of including time-based fixed effects.

Table 1, Column (1) reports an estimate of the effects of public revelation on the gap between the actual and predicted feedback scores. We can see that after the switch to revelation, the gap increased. The effect size of 0.13 is about 8% of the population standard deviation in Δs . All standard errors in this table are clustered at the level of the individual employer.

Table 1: Effects of “private feedback” public revelation on aggregate private feedback scores

	<i>Dependent variable:</i>			
	Δs , (Actual - Predicted) Private FB Ratings			
	(1)	(2)	(3)	(4)
Post-revelation	0.133*** (0.007)	0.167*** (0.007)	0.138*** (0.013)	0.067*** (0.018)
Post \times Month				0.011*** (0.002)
Constant	-0.013* (0.007)	-0.105*** (0.006)		
Less than 25 contracts for C and E	N	Y	Y	Y
Employer FE	N	N	Y	Y
Observations	899,842	537,640	537,640	537,640
R ²	0.001	0.002	0.498	0.498
Adjusted R ²	0.001	0.002	0.184	0.184

Notes: This table reports regressions where the outcome is the monthly aggregate feedback and the predictor is an indicator variable for the private feedback revelation. Both specifications control for month-specific effects, with Column (1) utilizing a fixed effects model, and Column (2) a random effects model. Significance indicators: $p \leq 0.05$: *, $p \leq 0.01$: **, and $p \leq .001$: ***.

One limitation of including all assignments as the unit of analysis is that it over-weights employers and workers with many contracts. In Column (2), we restrict the sample to employers and workers with fewer than 25 completed contracts in total. We can see that the effect size is somewhat larger with this restricted sample, but is broadly similar in magnitude.

As we noted in Section 3, one reason why a gap might emerge between some measure and an alternative measure is that in the post-period is selection of raters with idiosyncratically large or small gaps. To assess this possibility, in Column (3) we add an employer-specific fixed effect to the regression. The effect size is somewhat smaller when the fixed effects are included, but the coefficient implies that the increase in the gap between the private numerical rating *within* employers is quite close to the average effect.

As we saw in Figure 7, there was visual evidence for a change in the trend and not just a level difference. As such, for our preferred specification, in Column (4), we include both a

post-indicator and a linear time trend for the post period and continue using the restricted sample and the employer-specific fixed effect. We can see that some of the treatment effect detected in Columns (1)-(3) was the accumulation of a trend of an increasing gap in the post-period.

6 Conclusion

This paper documents that the reputation system in an online marketplace was subject to inflation—we observe systematically higher scores over time, which cannot be fully explained by improvements in fundamentals. Data from four other marketplaces exhibit the same trend, suggesting that reputation inflation is widespread. We develop an approach to quantify inflation based on using alternative measures of rater satisfaction. A market intervention that increased the the costs of negative feedback by making previously private feedback public, yielded data supporting the role of reflected costs as the root cause of inflation.

A strong rate of inflation—even if episodic and then contained—causes individual reputations to vary based on when a score was assigned, creating unwanted variation that undermines the usefulness of a system. Furthermore, if the pooling of feedback scores in the highest feedback “bin” becomes acute, statistical corrections cannot recover the lost information. An interesting next step is to explore the consequences of reputation inflation. In Appendix E, we present evidence that feedback scores have become less informative over time in our setting.

Reputation inflation is likely most acute in peer-to-peer platforms, such as online labor and sharing economy markets, where wage penalties for workers and employers’ reflected costs are high. Reasons for the higher worker cost of negative feedback include that feedback scores are often the sole signal of quality, workers are typically highly substitutable and have few transactions, and hence each rating is more consequential. As transactions are more personal, the reflected costs for employers are also likely higher. In contrast, when individuals assign feedback to products (e.g. movie reviews) there is likely no reflected cost, and inflation will be less acute. Indeed, numerical scores on such platforms are characterized by lower averages, a higher spread, and, in some cases, a decreasing temporal pattern (Cabral and Hortaçsu, 2010; Moe and Schweidel, 2012; Godes and Silva, 2012).

Reputation inflation is seemingly also present in the non-digital world. For example, there is widespread concern about grade inflation, and some schools have taken steps to counter it (Butcher et al., 2014). The debate found in this literature mirrors many of the issues we examine in this paper, namely whether the increase in grades is due to fundamentals or due to lower rating standards, and whether inflated ratings reduce grade informativeness and reduce student effort (Babcock, 2010).

For would-be marketplace designers, our paper illustrates a core market design prob-

lem, and elucidates its root cause. Whether there are effective platform design responses to this phenomenon is an open question. Changes in the reputation system, such as adding a higher ceiling in the feedback scores or additional dimensions of reputation, may temporarily mitigate—but do not solve—the problem.¹⁴ Platforms could provide monetary incentives for users who generate feedback, as dissatisfied users often leave no feedback (Nosko and Tadelis, 2015), but this approach could be costly to implement. Platforms could also emphasize reviewers as performing a service for fellow consumers, or provide other incentives for honest reviews: Yelp employs mechanisms such as badges for top reviewers, and makes the feedback score distribution of each reviewer publicly accessible. Mandatory grading curves are often employed in non-digital reputation systems, although it is challenging to force a distribution in settings where buyers evaluate sellers as a “flow.”¹⁵

Platforms already take steps to lower raters’ costs. These steps, such as simultaneously-revealed ratings (in place since the start of the platform) and anonymizing ratings through aggregation (as was the case with the private feedback change), did not prevent inflation from occurring in our data. Our analysis of the private feedback revelation quasi-experiment supports that raters’ costs are to a large part due to raters incurring a greater personal cost—or guilt—the greater the harm they impose on the rated worker.

An interesting alternative explanation would be that, once private feedback is revealed, workers are more likely to let the employers know that private feedback ratings have become consequential. In turn, employers may feel more pressure to leave higher private feedback ratings, or may become better informed about how harmful bad private feedback is. An interesting next step is to disentangle the effects of these channels.

¹⁴See also <https://www.youtube.com/watch?v=K005S4vxi0o>.

¹⁵Officer evaluation reports in the US Army limit senior raters to indicating only 50% or less of the officers they rate as “most qualified.”

References

- Abeler, Johannes, Daniele Nosenzo, and Collin Raymond**, “Preferences for truth-telling,” *Econometrica*, 2019, *87* (4), 1115–1153.
- Aperjis, Christina and Ramesh Johari**, “Optimal windows for aggregating ratings in electronic marketplaces,” *Management Science*, 2010, *56* (5), 864–880.
- Athey, Susan, Juan Camilo Castillo, and Bharat Chandar**, “Service quality in the gig Economy: Empirical evidence about driving quality at Uber,” *Available at SSRN*, 2019.
- Avery, Christopher, Paul Resnick, and Richard Zeckhauser**, “The market for evaluations,” *The American Economic Review*, 1999, *89* (3), pp. 564–584.
- Babcock, Philip**, “Real costs of nominal grade inflation?: New evidence from student course evaluations,” *Economic Inquiry*, 2010, *48* (4), 983–996.
- Benson, Alan, Aaron Sojourner, and Akhmed Umyarov**, “Can reputation discipline the gig economy? Experimental evidence from an online labor market,” *Management Science*, 2019.
- Berentsen, Aleksander, Guido Menzio, and Randall Wright**, “Inflation and unemployment in the long run,” *American Economic Review*, 2011, *101* (1), 371–98.
- Bolton, Gary, Ben Greiner, and Axel Ockenfels**, “Engineering trust: Reciprocity in the production of reputation information,” *Management Science*, 2013, *59* (2), 265–285.
- Büschken, Joachim, Thomas Otter, and Greg M Allenby**, “The dimensionality of customer satisfaction survey responses and implications for driver analysis,” *Marketing Science*, 2013, *32* (4), 533–553.
- Butcher, Kristin F, Patrick J McEwan, and Akila Weerapana**, “The effects of an anti-grade-inflation policy at Wellesley College,” *The Journal of Economic Perspectives*, 2014, *28* (3), 189–204.
- Cabral, Luis and Ali Hortaçsu**, “The dynamics of seller reputation: Evidence from eBay,” *The Journal of Industrial Economics*, 2010, *58* (1), 54–78.
- Chevalier, Judith A, Yaniv Dover, and Dina Mayzlin**, “Channels of impact: User reviews when quality is dynamic and managers respond,” *Marketing Science*, 2018, *37* (5), 688–709.
- Dellarocas, Chrysanthos**, “The digitization of word of mouth: Promise and challenges of online feedback mechanisms,” *Management Science*, 2003, *49* (10), 1407–1424.

- , “Reputation mechanism design in online trading environments with pure moral hazard,” *Information Systems Research*, 2005, *16* (2), 209–230.
- **and Charles A Wood**, “The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias,” *Management Science*, 2008, *54* (3), 460–476.
- Diewert, W Erwin**, “Index number issues in the consumer price index,” *Journal of Economic Perspectives*, 1998, *12* (1), 47–58.
- Einav, Liran, Chiara Farronato, and Jonathan Levin**, “Peer-to-peer markets,” *Annual Review of Economics*, 2016, *8*, 615–635.
- Filippas, Apostolos and John J Horton**, “The tragedy of your upstairs neighbors: Externalities of home-sharing,” Technical Report 2018.
- , – , **and Richard J Zeckhauser**, “Owning, using and renting: Some simple economics of the “sharing economy”,” *Management Science*, forthcoming.
- Fradkin, Andrey, Elena Grewal, and David Holtz**, “Reciprocity in Two-sided Reputation Systems: Evidence from an Experiment on Airbnb,” Technical Report, Working Paper 2019.
- Friedman, Milton**, “Nobel lecture: Inflation and unemployment,” *Journal of political economy*, 1977, *85* (3), 451–472.
- Ghose, Anindya, Panagiotis G Ipeirotis, and Beibei Li**, “Examining the impact of ranking on consumer behavior and search engine revenue,” *Management Science*, 2014, *60* (7), 1632–1654.
- Godes, David and José C Silva**, “Sequential and temporal dynamics of online opinion,” *Marketing Science*, 2012, *31* (3), 448–473.
- Hall, Jonathan V and Alan B Krueger**, “An analysis of the labor market for Ubers driver-partners in the United States,” *ILR Review*, 2018, *71* (3), 705–732.
- Horton, John J**, “Online labor markets,” *Internet and network economics*, 2010, pp. 515–522.
- , **David G Rand, and Richard J Zeckhauser**, “The online laboratory: Conducting experiments in a real labor market,” *Experimental Economics*, 2011, *14* (3), 399–425.
- Hu, Nan, Paul A Pavlou, and Jie Zhang**, “On self-selection biases in online product reviews,” *MIS Quarterly*, 2017, *41* (2).

- Huang, Guofang and K Sudhir**, “The Causal Effect of Service Satisfaction on Customer Loyalty,” *Available at SSRN 3391242*, 2019.
- Jin, Ginger Zhe and Phillip Leslie**, “The effect of information on product quality: Evidence from restaurant hygiene grade cards,” *The Quarterly Journal of Economics*, 2003, *118* (2), 409–451.
- Kamakura, Wagner A**, “Common methods bias,” *Wiley International Encyclopedia of Marketing*, 2010.
- Katz, Lawrence F and Alan B Krueger**, “The rise and nature of alternative work arrangements in the United States, 1995–2015,” *ILR Review*, 2019, *72* (2), 382–416.
- Klein, Tobias J, Christian Lambertz, and Konrad O Stahl**, “Market transparency, adverse selection, and moral hazard,” *Journal of Political Economy*, 2016, *124* (6), 1677–1713.
- Levin, Jonathan D**, “The Economics of Internet Markets,” Technical Report, National Bureau of Economic Research 2011.
- Li, Xinxin and Lorin M Hitt**, “Self-selection and information role of online product reviews,” *Information Systems Research*, 2008, *19* (4), 456–474.
- Liu, Qingmin**, “Information acquisition and reputation dynamics,” *The Review of Economic Studies*, 2011, *78* (4), 1400–1425.
- Luca, Michael**, “Reviews, reputation, and revenue: The case of Yelp.com,” *Working Paper*, 2016.
- and Georgios Zervas**, “Fake it till you make it: Reputation, competition, and Yelp review fraud,” *Management Science*, 2016, *62* (12), 3412–3427.
- Mayzlin, Dina, Yaniv Dover, and Judith Chevalier**, “Promotional reviews: An empirical investigation of online review manipulation,” *The American Economic Review*, 2014, *104* (8), 2421–55.
- Mishkin, Frederic S**, “Inflation targeting in emerging-market countries,” *American Economic Review*, 2000, *90* (2), 105–109.
- Moe, Wendy W and David A Schweidel**, “Online product opinions: Incidence, evaluation, and evolution,” *Marketing Science*, 2012, *31* (3), 372–386.
- Nosko, Chris and Steven Tadelis**, “The limits of reputation in platform markets: An empirical analysis and field experiment,” Technical Report, National Bureau of Economic Research 2015.

- Podsakoff, Philip M, Scott B MacKenzie, and Nathan P Podsakoff**, “Sources of method bias in social science research and recommendations on how to control it,” *Annual review of psychology*, 2012, *63*, 539–569.
- Proserpio, Davide and Georgios Zervas**, “Online reputation management: Estimating the impact of management responses on consumer reviews,” *Marketing Science*, 2017, *36* (5), 645–665.
- Resnick, Paul, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman**, “Reputation systems,” *Communications of the ACM*, 2000, *43* (12), 45–48.
- , **Richard Zeckhauser, John Swanson, and Kate Lockwood**, “The value of reputation on eBay: A controlled experiment,” *Experimental Economics*, 2006, *9* (2), 79–101.
- Sidrauski, Miguel**, “Inflation and economic growth,” *Journal of Political Economy*, 1967, *75* (6), 796–810.
- Stanton, Christopher T and Catherine Thomas**, “Landing the first job: The value of intermediaries in online hiring,” *The Review of Economic Studies*, 2015, *83* (2), 810–854.
- Sundararajan, Arun**, “From Zipcar to the sharing economy,” *Harvard Business Review*, 2013, *1*.
- Tadelis, Steven**, “Reputation and feedback systems in online platform markets,” *Annual Review of Economics*, 2016, *8*, 321–340.
- and **Florian Zettelmeyer**, “Information disclosure as a matching mechanism: Theory and evidence from a field experiment,” *The American Economic Review*, 2015, *105* (2), 886–905.
- Yoganarasimhan, Hema**, “The value of reputation in an online freelance marketplace,” *Marketing Science*, 2013, *32* (6), 860–891.

A Other reasons for the feedback score increase

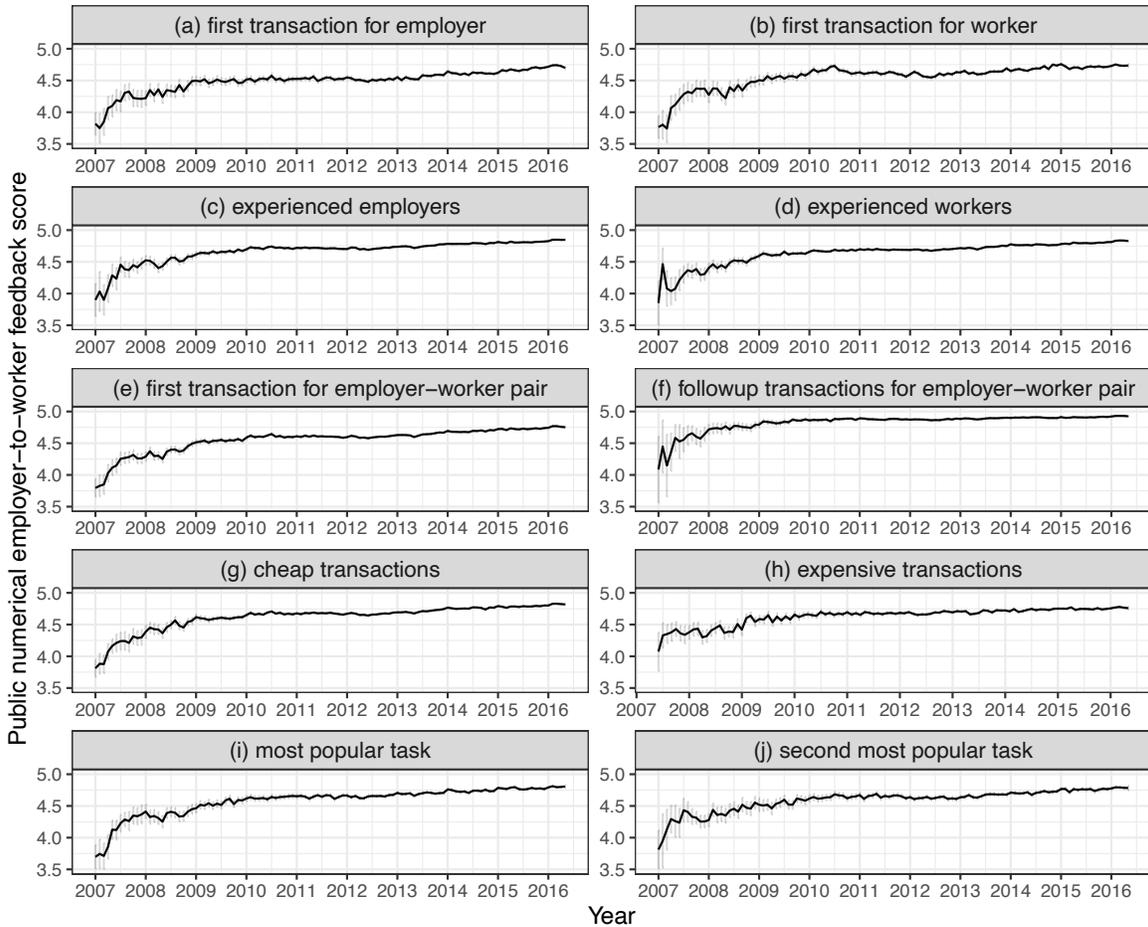
Figures 1 and 2 document an over-time increase in numerical feedback scores. This increase may be the outcome of changes related to employer/worker composition, experience, and selection, as well as changes in the composition of type of work that takes place on the platform. Disentangling these reasons from reputation inflation is the main purpose of our paper; we summarize our arguments below, and offer additional tests that directly rule out some plausible hypotheses.

Section 3 addresses these concerns through the alternative measures approach. This approach sidesteps the problem of estimating latent employer utilities, and instead measures how the gap between the primary and an alternative measure of employer satisfaction changes over time. Another advantage of this approach is that, insofar the alternative measure of satisfaction is also subject to inflation, we underestimate the degree of reputation inflation, and hence we produce a conservative estimate. One potential problem with this approach is that there may be systematic changes in how raters translate their utilities to the primary and alternative measures. This hypothesis is not supported by our data, as (1) using private feedback as an alternative measure of employer satisfaction, we restrict our attention to a short time horizon where such changes are unlikely to have take place, and (2) using written feedback as an alternative measure, results, we find similar patterns and estimates of inflation regardless of the training period employed in learning the \hat{s} function. Another potential drawback of this approach is that learning the \hat{s} function may introduce an estimation error. There is no evidence that this is the case in our data, because (1) the \hat{s} function employed for translating private feedback to numerical feedback is very simple, and (2) several different \hat{s} functions for translating private feedback to numerical feedback yield qualitatively similar results; we provide further details in Appendices B and C.

While our view is that the alternative measures approach provides strong evidence supporting the reputation inflation story, we believe it is useful to also offer supplementary evidence directly ruling out plausible alternative hypothesis for the observed increase in feedback scores. Plausible alternative explanations include that better employer and worker cohorts are active on the platform, that employers continuously transact with a subset of desirable workers, and that job composition skews towards inherently higher-rated tasks.

Changes in employer and worker composition could result in different trends in the observed increase in feedback scores. In Figure 8, panels (a) and (b) we plot the average employer feedback scores for the first transactions of employers and workers respectively, whereas panels (c) and (d) depict the average feedback scores only for transactions in which employers and workers respectively are experienced—employers and workers with more than four previous transactions in the platform. Strikingly, the same trend persists across all subpopulations, suggesting that the observed increase is not a function of experience: in-

Figure 8: Monthly average feedback scores assigned by employers to workers for various subsets of the online labor market data.



Notes: This figure plots the average public feedback scores in our setting. Scores are assigned by employers to workers upon the completion of each transaction, and the scale for feedback is 1 to 5 stars. For each observation, average scores are computed for every time period, and a 95% interval is depicted for every point estimate. Panels (a) and (b) plot average scores for the first employers’ and workers’ transactions respectively. Panels (c) and (d) plot transactions where employers and workers respectively, had more than 4 previous transactions on the platform. Panels (e) and (f) plot average scores for each employer-worker pair’s first, and followup transactions respectively. Panels (g) and (h) plot average scores for transactions with cost of less than 100 and more than 1000 U.S. dollars respectively, and panels (i) and (j) plot average scores for the two most common freelancing tasks. We obtain quantitatively similar results when we choose other thresholds to define “experience,” when we choose other thresholds to define “cheap” and “expensive” transactions, and when we restrict our sample to various other freelancing tasks.

experienced employers give higher ratings over time, and inexperienced workers also receive higher ratings. The average scores of experienced users are only slightly higher than those of first-time users, suggesting that some selection takes place, but subpopulations exhibit the

same over-time increase.

The observed increase in feedback scores could also be explained by trading partner selection, that is, employers identifying a subset of desirable workers, and continuously transacting with them. In Figure 8, panel (e) plots average feedback scores only from first-time transactions between an employer-worker pair, whereas panel (f) plots average feedback scores from follow-up transactions—transactions where the employer-worker pair has transacted in the past. When an employer chose to transact with the same worker, we expect that the employer was more satisfied. Indeed, we observe that average scores start higher, yet we see a similar over-time increase in average numerical scores across both subpopulations.

An additional piece of evidence against this type of selection comes from the observed trends in the marketplaces depicted in Figure 2. In the home-sharing marketplace, it is unlikely that selection is a major factor, because users are unlikely to travel to the same destination for leisure repeatedly. In the service marketplace, the platform matches providers with consumers, and hence selection is ruled out by default. Furthermore, provider/worker capacity is highly constrained in all marketplaces, making it unlikely that the same provider will be available in the future. Together, this empirical evidence supports that selection cannot explain the observed increase.

Changes in job type composition could also explain the observed trends—for example, the type of transactions taking place in our setting could skew towards an inherently highly-rated transaction type over time. In Figure 8, panel (g) plots average feedback scores for transactions worth less than 100 US dollars, panel (h) plots average scores for transactions worth more than 1000 dollars, and panels (i) and (j) plot average scores for the two most frequent types of tasks. The feedback scores for these job subpopulations exhibit a strikingly similar pattern, supporting the view that the composition of types of transaction in the platform also cannot explain the observed increase. Furthermore, an additional piece of evidence comes from the fact that the types of products and services transacted in the marketplaces depicted in Figure 2 remain constant throughout our data collection period.

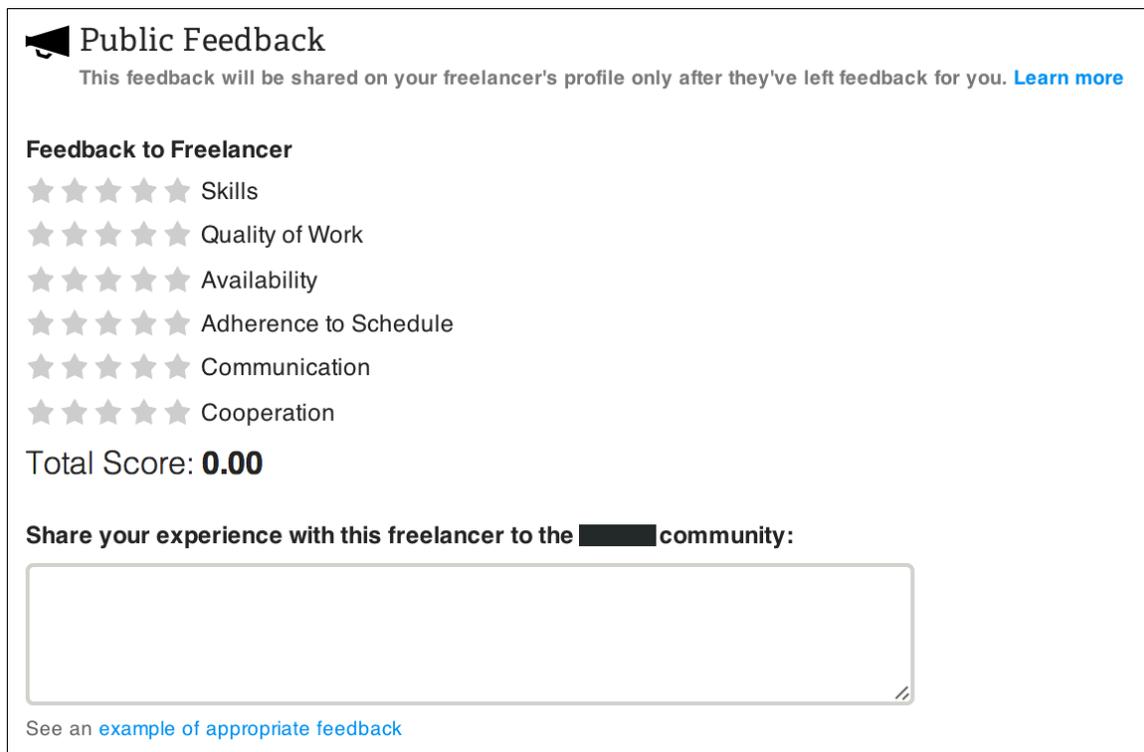
B Robustness tests for private feedback

B.1 Details on private feedback elicitation

The status-quo employer-on-worker feedback is shown in Figure 9a. For private feedback solicitation, the interface shown in Figure 9b was simply appended at the end of the public feedback form. As such, employers had to assign both public and private feedback for the same transactions.

Figure 9: Public and private employer-on-worker feedback interfaces.

(a) Public feedback interface.



Public Feedback
This feedback will be shared on your freelancer's profile only after they've left feedback for you. [Learn more](#)

Feedback to Freelancer

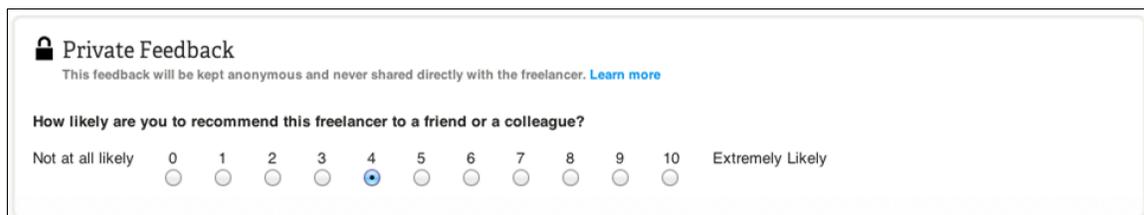
- ★★★★★ Skills
- ★★★★★ Quality of Work
- ★★★★★ Availability
- ★★★★★ Adherence to Schedule
- ★★★★★ Communication
- ★★★★★ Cooperation

Total Score: **0.00**

Share your experience with this freelancer to the [redacted] community:

See an [example of appropriate feedback](#)

(b) Private feedback interface.



Private Feedback
This feedback will be kept anonymous and never shared directly with the freelancer. [Learn more](#)

How likely are you to recommend this freelancer to a friend or a colleague?

Not at all likely 0 1 2 3 4 5 6 7 8 9 10 Extremely Likely

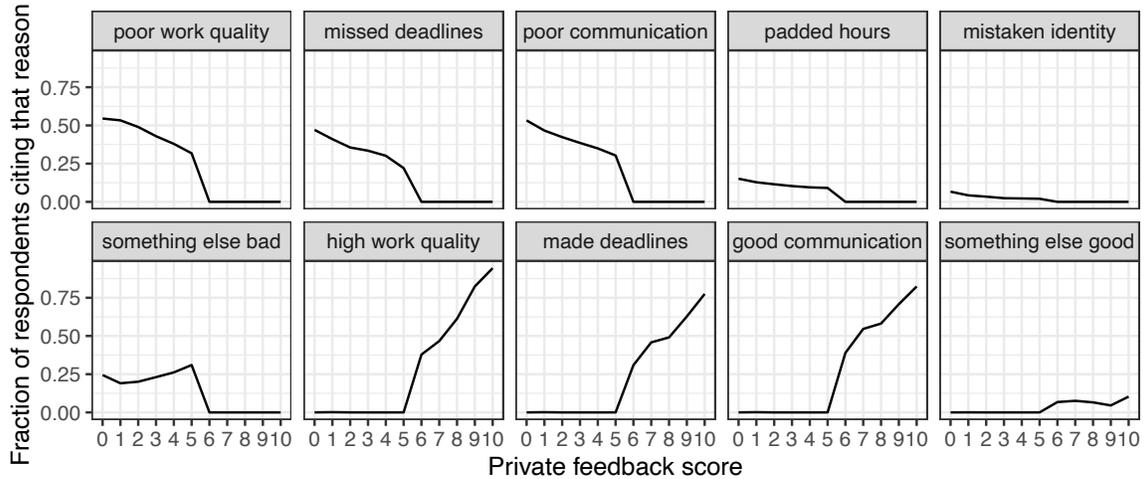
One concern is that the elicitation of private feedback information allowed employers who

had a negative experience to assign workers a bad private feedback score instead of a bad public feedback score. Had this been the case, we would expect to see either a discontinuity in public feedback score averages, or a change in the rate of their increase. We see no such pattern in Figure 1b. Similarly, in Figure 4b, we see no such pattern in the sentiment expressed in written employer feedback.

B.2 Misinterpreting private feedback

One concern with any new feedback feature is that raters might simply not understand the new ratings. However, we have evidence that employers, at least collectively, understood quite well what the scale meant. When asked for private feedback, the platform also displayed a set of reasons that the employer could optionally select to indicate the reason for their score. Positive reasons were shown when the assigned feedback was above 5, while negative reasons were shown otherwise (during the 0 to 10 scale period). We use this “reason” information to verify that employers did not misinterpret the private feedback question. The fractions of private feedback reports citing these different reasons against the assigned private feedback score (1 to 10 scale) are plotted in Figure 10. We can see that there is a clear trend in the “correct” direction for both scores, indicating that private feedback scores were correctly assigned, at least on average.

Figure 10: Fraction of users citing a given reason when giving private feedback, by score.



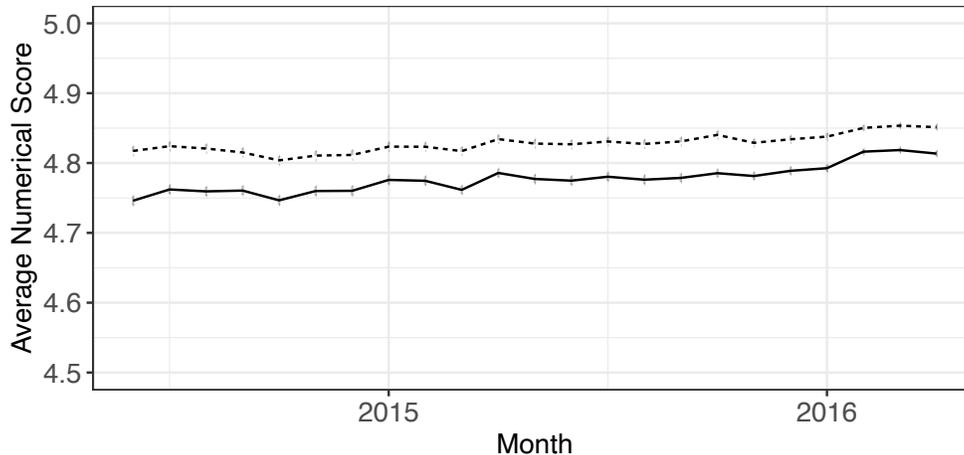
Notes: This figure plots the fraction of feedback reports that cited each reason as the basis of the feedback being positive or negative, against the private feedback score given. Across every case, we notice that employers that assigned more extreme feedback scores tend to cite reasons of the same sentiment more frequently.

B.3 Selection issues

Another plausible concern is that employers could be self-selecting into when they will leave private feedback, and that changes in private feedback scores reflect changes in the selection process. Figure 11 plots the evolution of numerical public feedback for all contracts (solid line), and contracts for which private feedback was also assigned (dashed line). We observe that contracts in which private feedback is also assigned receive higher average public ratings, implying that employer who publicly indicate higher satisfaction are more likely to also assign private feedback. The two lines closely resemble each other throughout the period where we collect both types of feedback, indicating no systematic change over time.

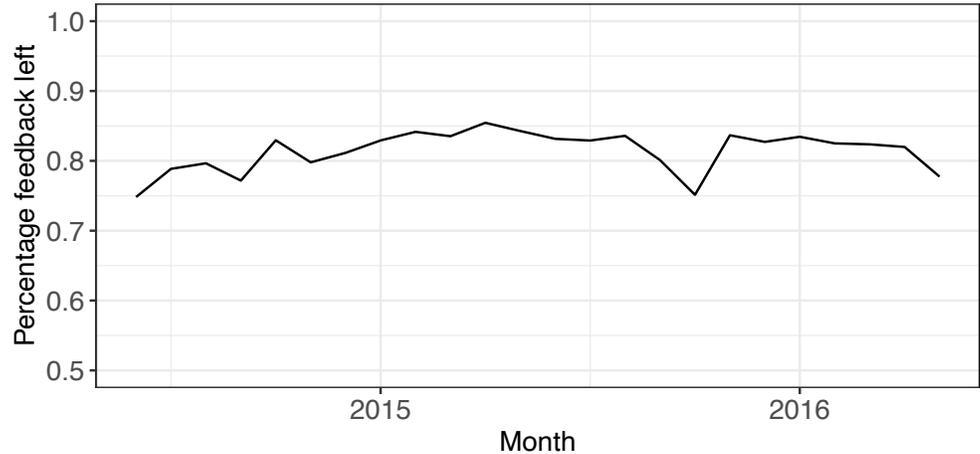
Another concern is that employers decision to leave private feedback when they leave public feedback could change over time. Figure 12 plots the percentage of contracts that received private feedback amongst these contracts that received public feedback. We observe that there is no systematic change over time in employers' decisions to assign private feedback when they assign public feedback. Furthermore, the percentage of employers that chooses to leave private feedback is high, with an average of 81.4% of employers deciding to also assign private feedback.

Figure 11: Average public numerical scores for all contracts, and for contracts to which private feedback was assigned.



Notes: This figure plots the monthly average feedback scores for all contracts (solid line), and the monthly average feedback scores for contracts for which private feedback was also assigned. A 95% interval is depicted for every observation. Scores are assigned upon the completion of each transaction, and the scale for numerical feedback is 1 to 5 stars.

Figure 12: Percentage of employers leaving private feedback in addition to public numerical feedback.



Notes: This figure plots the monthly percentage of contracts for which employers assigned private feedback, amongst those contracts for which employers also assigned numerical feedback.

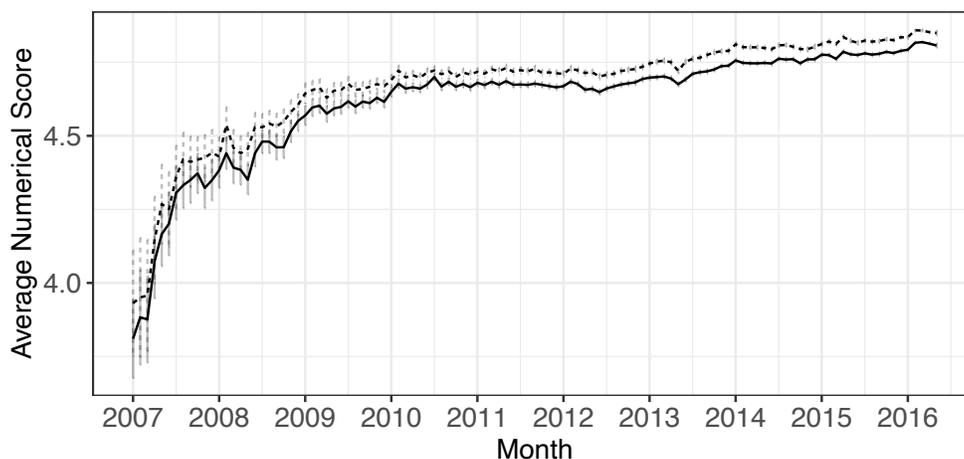
C Robustness tests for written feedback

C.1 Selection issues

A concern about the use of written feedback as an alternative measure of rater satisfaction is that employers' assignment behavior changes over time. In what follows we conduct robustness tests to identify potential sources of bias for our analysis.

As with private feedback, a plausible concern is that employers may be more or less satisfied when deciding to assign written feedback in addition to numerical feedback. Figure 13 plots the evolution of numerical feedback for all contracts (solid line), and all contracts for which written feedback was also assigned (dashed line). We observe that contracts in which written feedback is also assigned receive higher ratings, implying that more satisfied employers assign written feedback. However, the degree to which this bias occurs does not change throughout our data. Furthermore, since written feedback is positively biased, comparing the predicted scores from text versus the evolution of all scores gives us a lower bound for the degree of inflation.

Figure 13: Monthly average numerical scores, and monthly average numerical scores when written feedback was assigned.

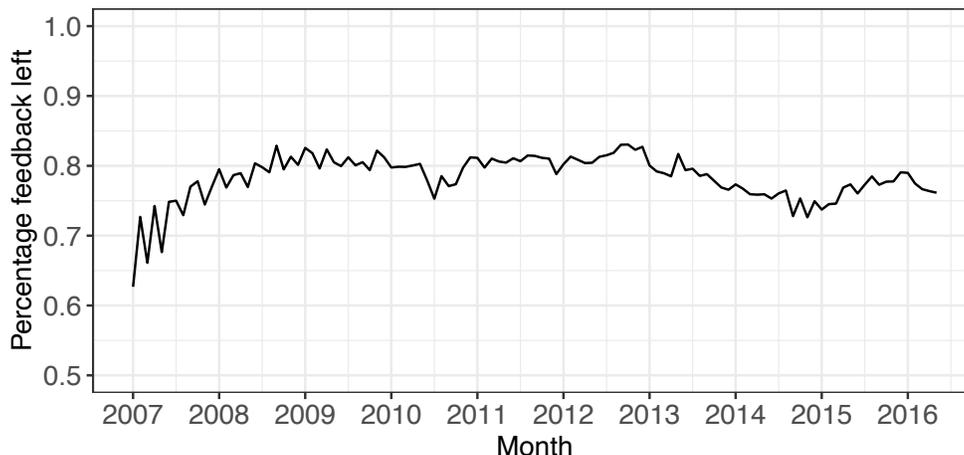


Notes: This figure plots the monthly average feedback scores for all contracts (solid line), and the monthly average feedback scores for contracts to which written feedback was also assigned (dashed line). Scores are assigned upon the completion of each transaction, and the scale for feedback is 1 to 5 stars. For each observation, average scores are computed for every time period, and a 95% interval is depicted for every point estimate.

Similarly to private feedback, a concern is that employers decision to leave written feedback when they leave public feedback could change over time. Figure 14 plots the percentage of contracts that received written feedback for those contracts that also received public feedback. We observe that there is no systematic change over time in employers' decisions to

assign private feedback when they assign public feedback. The percentage of employers that chooses to leave written feedback is also high, with an average of 79.2% of employers deciding to also assign written feedback.

Figure 14: Percentage of employers leaving written feedback in addition to public numerical feedback.



Notes: This figure plots the monthly percentage of contracts for which employers assigned written feedback, amongst those contracts for which employers also assigned numerical feedback.

C.2 Composition of raters

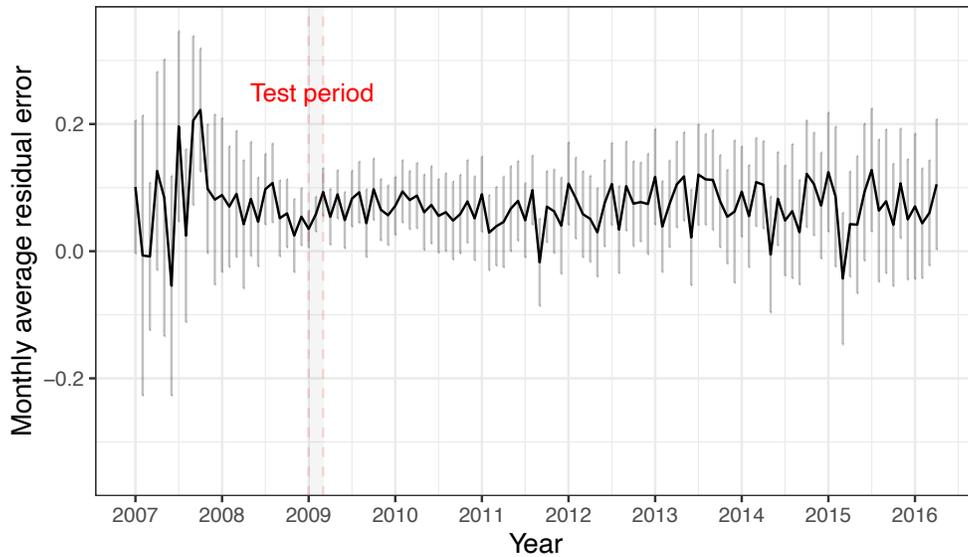
Shifts in the composition of raters could potentially introduce bias in using written feedback as an alternative measure of satisfaction. More specifically, the widening gap between numerical scores and scores predicted from written feedback could be the outcome of employers with this rating behavior—employers who assign higher scores for the same written feedback—joining the platform over time, or, equivalently, employers with the opposite rating behavior dropping out. In the language introduced in Section 4, this issue can be thought of as a systematic changes in $\mathbb{E}_{F'}[\epsilon]$.

We test against this hypothesis as follows. For a period of time T , we compute the average residual error r_i for each employer i that left feedback during T , defined as the divergence between the numerical scores and the predicted scores from the associated written feedback employer i assigned. The employer average residual error is then $\bar{r}_T = \sum_{i \text{ left feedback in } T} r_i$. We then test whether, amongst these employers, there is a systematic drop-out behavior that has led to employers with wider gaps remaining in the platform in the post period (and, respectively, whether only employers with narrower gaps were present in the pre-period). We can do so by simply computing $\bar{r}_t = \sum_{i \text{ left feedback in } T \text{ and } t} r_i$, for any $t \neq T$. If for $t > T$

the quantities \bar{r}_t show a systematic increase, then this composition shift in rater types may bias our estimates.

Figure 15 carries out this analysis for employers who left feedback in January and February of 2009. For the predicted scores, we employ the predictions of the model in the lower panel of Figure 16. We find no evidence of a systematic trend in neither the pre-period, nor the post-period, suggesting that our inflation estimates are not subject to this source of bias. Conducting the analysis for other periods in our data or for other predictive models, yields qualitatively identical results.

Figure 15: Employer average residual error in for employers who left feedback during January and February 2009.



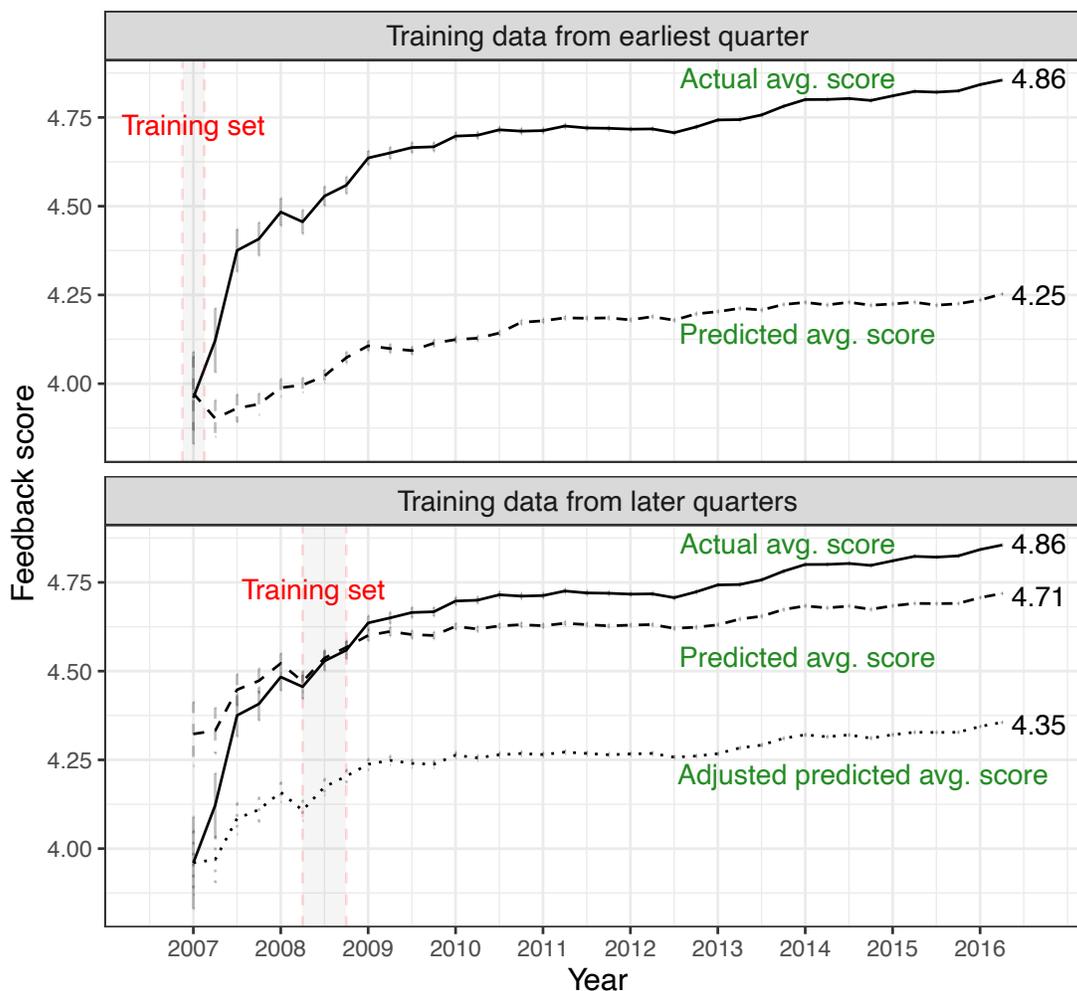
Notes: This figure plots the employer average residual error over time for the set of employers who left feedback during the period indicated by the shaded area. The average residual errors are computed for every month, and a 95% interval is depicted for every point estimate.

C.3 Alternative training periods

In the bottom panel of Figure 16 we perform the same empirical exercise as in Section 3.2, again plotting the average quarterly feedback over time, for both the numerical public feedback and the feedback predicted from the written feedback. However, our training sample now comes from a longer time period indicated by the two vertical red lines, and is larger, consisting of 10,555 feedback samples. As expected, the predicted and actual scores closely match up during the training period. However, in the period before, the predicted score is higher than the numerical score, and the opposite holds after the training period. We adjust the second score by a constant, so that the predicted score matches the actual feedback score

in the beginning of our data. With this adjustment, the average predicted feedback score at the end of the data “should” have only been 4.35 stars. Using the first quarter sample, the point estimate is that 67.7% of the increase in feedback scores is due to inflation, whereas the larger sample from the middle of the data implies 56.6% of the increase is due to inflation. Reassuringly, the two corpora give similar results.

Figure 16: Numerical public feedback and predicted score from textual feedback using the first quarter as the training period.



Notes: This figure plots the evolution of average public feedback scores (solid line) versus the average predicted score of textual feedback (dashed line) assigned by employers to workers. A 95% interval is depicted for every point estimate. The shaded area indicates the quarters from which training data was obtained for the corresponding predictive model. The training sets consist of 1,492 samples (top panel) and 10,555 samples (bottom panel). Adjusted predicted scores (dotted line in the bottom panel) are calculated by subtracting the constant from the predicted scores that allows the left endpoints of the adjusted and actual score lines to coincide.

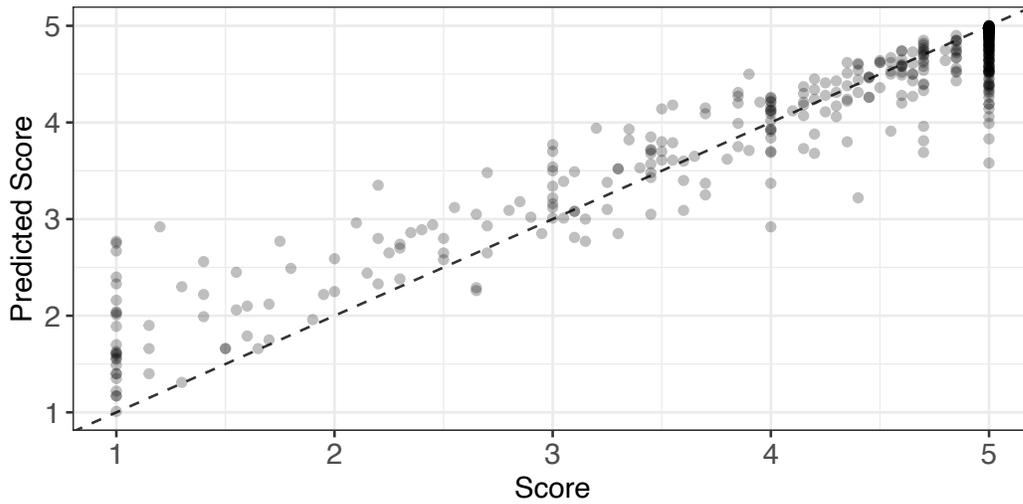
C.4 Predictive algorithm performance

We present more details about the performance of the algorithms used to extract the written feedback sentiment in Section 3.2.

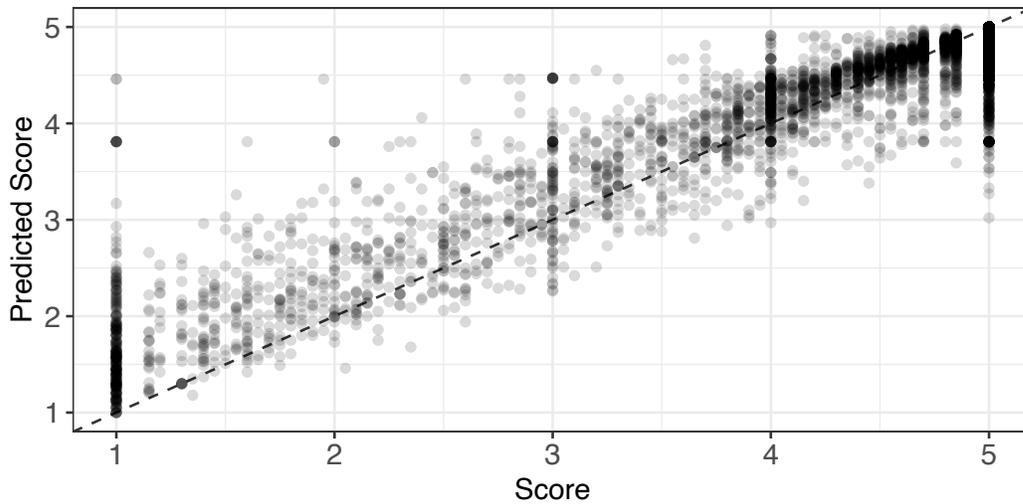
Figure 17a plots the scatterplot of numerical scores versus predicted scores from written feedback for the algorithm trained on data coming from the earliest quarter. Figure 17b plots the same scatterplot for the algorithm trained on data coming from the later quarters. Since the training data is skewed towards higher scores in both cases, the algorithms are expected to over-predict, but both predictive models attain good performance, with the mass of their predictions being close to the 45 degree line. Furthermore, note that this performance is attained despite the fact that we should expect somewhat large variance between scores and written feedback amongst different employers. The appropriateness and good performance of our models is further verified by the fact that the estimates we obtain closely match the performance of our model-free approach presented in Figure 5.

Figure 17: Numerical score versus predicted score from text scatterplot.

(a) Performance on training set from earliest quarter.



(b) Performance on training set from later quarter



Notes: The top panel plots the scatterplot of numerical scores assigned to contracts versus numerical scores predicted from the associated written feedback for the algorithm trained on data from the earliest quarter, while the bottom panel plots the same scatterplot for the algorithm trained on data from the later quarter. The scale for feedback is 1 to 5 stars. The 45 degree line represent the performance of a “perfect” prediction algorithm.

D Extensions to the model of Section 4

D.1 Convergence to equilibrium

We examine the equilibrium convergence process for the model developed in Section 4. We assume the following simple matching process: in every period, employers randomly match with workers, workers produce outputs, and employers subsequently report feedback. Let p_t denote the truth-telling fraction of the employer population after period t . To avoid cases where the convergence process is trivial, we will assume that the $p_E < 1$, that is, that the equilibrium truth-telling fraction is not the all-truthful equilibrium (see Equation 5).

Let the initial condition of the marketplace be $p_0 = 1$, that is, every employer starts off reporting feedback truthfully. Because employers and workers match randomly, a fraction $\theta_B = (1 - \theta)(1 - q_L) + \theta(1 - q_H)$ of the employers receives a bad output ($y = 0$) after every period. The employers who received a bad output then compare their benefit from truth-telling b with the cost of truthfully reporting bad feedback. Employers whose cost from truth-telling is lower than the benefit give bad feedback to the workers. As such, a fraction $l_0 = \theta_B[1 - F(\frac{b}{\Delta w(p_0)})]$ begins to lie after the first period, and hence $p_1 = p_0 - l_0$.

Let $T(x) = F(b/x)$ be the proportion of sellers that are better off truthfully reporting if the cost of bad feedback is x . From Equation 5 we obtain $T(p_E) = p_E$. Since F is a cumulative distribution function, and Δw is decreasing in its argument, T is a decreasing but non-negative function. As a result, $p_2 < p_1$, but $l_1 < l_0$, and hence $p_1 - p_2 < p_0 - p_1$. Following the same argument, we can inductively show that the dynamics of the marketplace result in convergence to the equilibrium truth-telling fraction p_E , and that the rate of convergence decreases as the market approaches the equilibrium point. This is the pattern that we observed empirically in all marketplaces: reputation initially inflates fast, but then flattens out as the equilibrium fraction is approached.

It is worth noting that the convergence process we examine assumes away other factors such as non-random matching, low-quality workers dropping out of the platform after receiving bad feedback, and so on. We assume these factors away not because we believe that they do not play an important role in the evolution of feedback scores, but rather because we attempt to examine the process of reputation inflation in isolation.

D.2 The impact of reputation inflation on the wage penalty

The model of Section 4 predicts that the wage penalty $\Delta w(p) = w_{s=1;p} - w_{s=0;p}$ is increasing in the truth-telling fraction p (see Equation 4). As reputation inflates, i.e., as p decreases, more employers give good feedback regardless of the output they receive, and hence good feedback becomes less informative. In the language of our model, the employers' posterior belief upon observing good feedback that a worker's quality is high, $\Pr(q = q_H | s = 1; p)$, is

increasing in p . However, the employers' posterior belief upon observing bad feedback that a worker's quality is high, $\Pr(q = q_H | s = 0; p)$, does not depend on p . To see why, note that for any p , employers who report truthfully will always assign bad feedback when they receive bad output, and hence the probability that a worker will receive bad feedback depends on the worker's quality q .

We next examine how alternative assumptions can change this result. Towards that end, we consider an extension to the model of Section 4 that departs from the base model in two ways. First, workers produce output $y \in \{-1, 0, 1\}$, where the new output $y = -1$ indicates a particularly bad outcome from which employers obtain utility equal to -1 . We assume that only low-quality workers can produce outcome $y = -1$, and that their production function is described by the probability distribution $\Pr(s = 1 | q = q_L) = q_L$, $\Pr(s = 0 | q = q_L) = 1 - q_L - \alpha$, and $\Pr(s = 1 | q = q_H) = \alpha$, with $\alpha < q_L$. Second, employers take worker output into account when assigning ratings, reporting truthful feedback if

$$b + |y| \geq c_i \Delta w.$$

In words, employers always report truthfully if they either receive output $y \in \{-1, 1\}$, and may choose to lie if they receive $y = 0$. All other model primitives remain the same. Note that the model of Section 4 is obtained by setting $\alpha = 0$.

The different cost structure changes the inferences that employers make upon observing the assigned feedback scores. When considering a particular worker that received bad feedback in the previous round, i.e., $s = 0$, the Bayesian employer now infers that

$$\Pr(q = q_H | s = 0; p) = \frac{(1 - q_H)p\theta}{(1 - q_H)p\theta + [\alpha + (1 - q_L - \alpha)p](1 - \theta)}.$$

Note that the p term no longer divides out. It is easy to see that $\Pr(q = q_H | s = 0; p)$ increases in p , and hence $w_{s=0;p}$ also increases in p , that is, bad feedback becomes costlier to receive as reputation inflates. To see why, consider the extreme case of full inflation, $p = 0$; then only low-quality workers are receive bad feedback when they produce outcome $y = -1$.

If the worker received good feedback, i.e., $s = 1$, the Bayesian employer infers that

$$\Pr(q = q_H | s = 1; p) = \frac{(q_H + (1 - q_H)(1 - p))\theta}{(q_H + (1 - q_H)(1 - p))\theta + (q_L + (1 - q_L - \alpha)(1 - p))(1 - \theta)}.$$

We can now show that $\Pr(q = q_H | s = 1; p)$ is increasing in p if $\frac{q_L}{q_H} \leq 1 - \alpha$, and is decreasing if the opposite condition holds. To see how reputation can make positive feedback more informative, consider the extreme case where $a = 1 - q_L$, and hence low-quality types never produce $y = 0$. Then reputation inflation only affects high-quality workers who can produce $y = 0$, and hence the wage premium $w_{s=1;p}$ increases as p decreases.

E Informational implications of reputation inflation

The impact of reputation inflation could be minimal if market participants “know” about the rate of inflation and adjust accordingly; even if individuals are not well-informed, the platform could implement statistical adjustments in its design of the reputation system to uncover the “true” (non-inflated) scores. However, if the pooling in the highest feedback “bin” becomes acute, statistical corrections cannot recover the lost information. This is partially due to the fact that, by design, numerical scale systems are prone to top-censoring; for the question “rate on a scale from 1 to X,” the value of X must be pre-specified.¹⁶

To see the problem created by top-censoring, consider the information conveyed by the observation of a binary variable X , as it is captured by the information-theoretic entropy $H(X) = p \log(p) + (1 - p) \log(1 - p)$, where p is the probability of one outcome. As p goes to either 1 or 0, the information conveyed by the variable—in our case, the observed feedback score—goes to zero. However, this binary characterization of the reputation system is a simplification that could elide an important way in which rising—and even more compressed scores—could convey just as much (or even more) information. Consider increasing all nominal scores by some fixed amount and then “shrinking” all scores toward some new higher mean. This transformation would have no informational implications. To assess informativeness, we need to take an empirical approach.

To assess the informativeness of the feedback scores about worker quality over time, we conduct a variance decomposition, showing how the fraction of unexplained variance in feedback scores changes over time. Suppose that the data generating process of a worker’s feedback is

$$\text{SCORE}_{it} = a_{it} + c_t + \epsilon_{it}, \tag{A1}$$

where a_{it} is the worker’s true quality, c_t is a baseline time effect, and ϵ_{it} is some noise term such that $E[\epsilon_{it}] = 0$.¹⁷ If, over time, more of the variation in feedback scores can be explained by the variation in the noise term rather than by variation in the quality of individuals, then a feedback score is becoming less informative of the worker’s true quality.

Consider a Bayesian employer trying to infer the quality of a worker from a score: the more the feedback score is attributable to noise, the lesser its impact on the employer’s posterior belief of worker’s quality after observing this score. To wit, let $\Pr(a) \sim N(a_0, \sigma_0^2)$ be the employer’s prior distribution for worker quality, and let $\epsilon \sim N(0, \sigma^2)$ be the noise term with known variance σ^2 , and a be the worker’s true quality, which the employer forms a posterior about after observing a feedback score. After observing the worker’s feedback score

¹⁶This is why reputation inflation differs from monetary inflation; a sandwich that used to cost \$0.50 and may now cost \$12. However, this could not happen if price was mechanically restricted to be below \$1.

¹⁷For simplicity, we are treating the feedback score as continuous. The logic is identical in the dichotomous case.

SCORE, the employer’s posterior is

$$\Pr(a|\text{SCORE}) = N\left(\frac{\frac{1}{\sigma^2}\text{SCORE} + \frac{1}{\sigma_0^2}a_0}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}\right).$$

From the above equation, as $\sigma^2 \rightarrow \infty$, $\Pr(a|\text{SCORE}) \rightarrow \Pr(a)$, or in words, as the noise component of the score explains more of the variance, the observed feedback becomes less informative, and at the limit, has no effect on the employer’s beliefs.

To explore the informativeness of feedback scores empirically, we make two assumptions. First, for a suitably small window of time (i.e., a quarter), we assume that the baseline time effect, c_t , is fixed. Second, we assume that the population distribution of a_{it} can have a changing mean, reflecting shifts in worker quality, but its variance is constant; workers could be getting systematically better or worse, but their abilities are not getting more or less spread out.

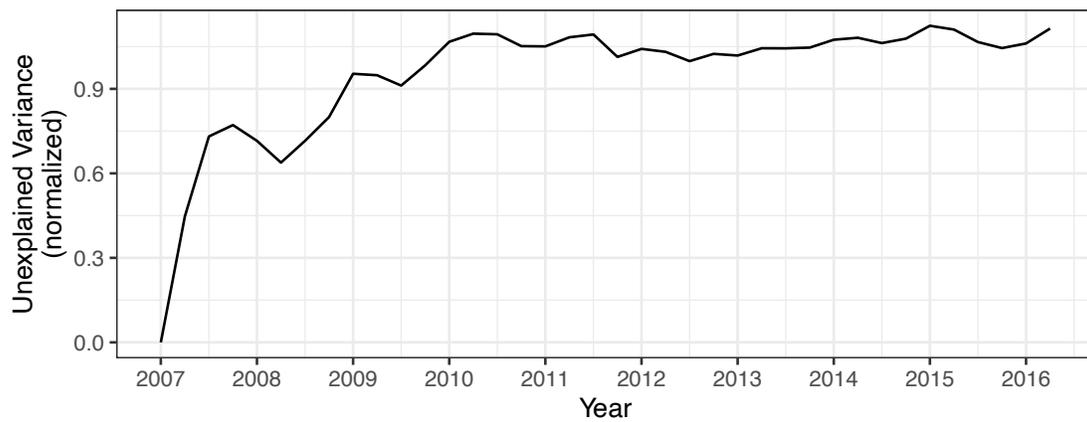
The fraction of variance due to noise is the quantity

$$\frac{\text{Var}(\epsilon)}{\text{Var}(\text{SCORE})} = 1 - \frac{\text{Var}(a)}{\text{Var}(\text{SCORE})}. \tag{A2}$$

If this ratio increases over time, feedback scores are becoming less informative. We can easily compute this fraction for a time period t by performing the regression implied by Equation A1—the quantity of Equation A2 is $1 - R_t^2$, where R_t^2 is the coefficient of determination from the period t regression.

We fit the regression described in Equation A1 on the feedback scores generated in every quarter of our data separately. On each of these regressions, we are using fixed worker effects to estimate a_{it} , thereby allowing worker quality to evolve in time, even “within” a worker. Figure 18 plots the percentage difference of $1 - R_t^2$ from the minimum unexplained variance, which is found at the first period in our data. The increase in unexplained variance from 2007 to 2016 is about 118% (from 0.32 to 0.70). This strong positive trend in the explained variance implies that the relative importance of noise in explaining feedback grows over time, which in turn implies that the informativeness of feedback about worker quality has deteriorated.

Figure 18: Feedback score variance not explained by worker quality over time. Scores are reported as percentage differences with respect to the minimum unexplained variance.



Notes: Unexplained variance is reported the percentage difference with respect to the minimum unexplained variance of the time series, which is attained at the first period of this figure. The data of each quarter consists of workers with at least 2 jobs in that quarter, as otherwise the fixed effect a_{it} would perfectly predict their feedback score. Utilizing different cutoffs does not quantitatively change our results.