Reputation Inflation[*]

Apostolos Filippas
Fordham

John J. Horton
MIT & NBER

Joseph M. Golden
Collage.com

October 3, 2021

Abstract

We show that average buyer ratings of sellers have grown substantially more positive over time in five online marketplaces. Although this increase could by explained by (i) marketplace improvements that increased rater satisfaction, it could also be caused by (ii) "reputation inflation," with raters giving higher ratings without being more satisfied. We present a method to decompose the growth in average ratings into components attributable to these two reasons. Using this method in one marketplace where we have extensive transaction-level data, we find that much of the observed increase in ratings is attributable to reputation inflation. We discuss the negative informational implications of reputation inflation and consider the likely causes.

[*] 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 well as some of their limitations (Dellarocas and Wood, 2008; Aral, 2014; Tadelis and Zettelmeyer, 2015; Tadelis, 2016; Hu et al., 2017). 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 these feedback scores in various online marketplaces seems implausibly rosy. 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. We show in this paper that 85% of rated workers in an online labor market—which we call our focal marketplace—received a perfect rating in recent years.[1] However, we also show that feedback scores did not start out this positively skewed: the fraction of workers receiving a perfect 5-star rating grew from 33% to 85% in just 6 years. Increasing average feedback scores in marketplaces seem to be commonplace: we collected data from five different online marketplaces, and each exhibits a marked increase in average feedback scores over time.

Rising feedback scores can be caused by two distinct—but not mutually exclusive—reasons: (1) raters are becoming more satisfied, or (2) raters are rating higher, despite not being more satisfied. 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 (or low-quality buyers/sellers exiting the platform), and lower-priced products. Improvements are obviously welcome, but with a fixed rating scale, they can lead to pooling of feedback scores at the highest possible score. The second possibility—raters giving higher scores despite not being more satisfied—can be described as a kind of inflation. This inflation can also cause pooling at the highest score, but is more worrisome because of its greater potential to reduce the informativeness and stability

[1] We use the terms "employer" and "worker" for consistency with the literature, and not as a comment on the legal relationship of the transacting parties.

of the reputation system.

If reputation inflation can explain at least some of the increase in average feedback scores, then we would expect a gap to emerge between what raters rate and how they actually feel about a transaction. To explore this possibility, we use information obtained by the introduction of a parallel and experimental reputation system into our focal marketplace. More specifically, a new feedback question 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.

We show that raters were far more candid in "private," with substantial numbers reporting dissatisfaction privately but still assigning a perfect 5-star rating publicly. This gap suggests that raters are reluctant to give negative feedback publicly because they do not want to harm the ratees' future prospects, either for altruistic reasons or because they fear retaliation of some kind. We also show 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 divergence provides some evidence of reputation inflation on the platform, albeit over a short time window when the private feedback question was asked.

To determine how much of the increase in average ratings is attributable to reputation inflation, we introduce a method for decomposing average ratings increases into the component that can be explained by changes in satisfaction and the component which cannot. To obtain a point estimate of the effect of inflation, the method requires an alternative measure of rater satisfaction not prone to inflation. Importantly, if the alternative measure is also prone to inflation, the method yields a lower bound. The method consists of learning the expected value for the actual numerical feedback, conditional upon the alternative measure of rater satisfaction. Under some mild assumptions likely to be met in practice, this learned conditional expectation function can then be applied to new transaction data, predicting what the average score "should" be given the alternative feedback data. This allows one to net out the increase not attributable to changes in marketplace fundamentals.

Alternative measures of rater satisfaction might seem hard to come by, but one measure available in many online marketplaces is the textual feedback that accompanies scores on the same transactions. For reasons we will discuss, textual feedback may be less prone to inflationary pressures. Consistent with this view, we show in our focal marketplace that the same sentences systematically have higher associated numerical feedback scores as time passes. 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.

Using written feedback as our alternative measure of rater satisfaction, we fit a model that predicts numerical feedback from the text of written feedback. We find that more than 50% of the increase in scores over a 6 year period was due to inflation, with this result being robust across different specifications and training sets. Insofar as written feedback is also subject to inflation, our approach *understates* the extent of reputation inflation. As numerical ratings are often accompanied by written feedback, this method can be readily used in other contexts.

A natural question is whether the reputation inflation we identify in our focal marketplace—and likely occurs in other marketplaces—"matters" in practice for the functioning of the reputation system. Although we do not explore this question empirically, there are several theoretical reasons why strong inflation in a system with a fixed, top-censored scale will cause a loss of information, analogous to the problems with grade inflation (Babcock, 2010; Butcher et al., 2014).

Our key contribution is documenting the extent of reputation inflation in a large online marketplace, by using an approach that accounts for changes in platform fundamentals. Our long-run, whole-system perspective is possible because we use data spanning over a decade of the operations of the marketplace. Although we cannot perform the same decomposition in other marketplaces, we observe increasing average feedback scores in every marketplace for which we could obtain data, even though none of these marketplaces allows "tit-for-tat" rating behavior (Bolton et al., 2013). Given that many online marketplaces share the same features as our focal marketplace, this evidence suggests that the problem is widespread.

The rest of the paper is organized as follows. Section 2 introduces the empirical setting, and documents increasing feedback scores over time across five online marketplaces. Section 3 presents descriptive evidence on the problem in our focal marketplace. Section 4 introduces our decomposition method and applies it. Section 5 discusses the causes and implications of reputation inflation. Section 6 concludes.

2 Empirical context and descriptive evidence of rising average feedback in several online marketplaces

Our focal market is a large online labor market (Horton, 2010). 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 (Filippas et al., 2020). Online markets offer a convenient setting for research due to the excellent measurement afforded in the online setting (Horton et al., 2011; Horton and Tambe, 2015).

2.1 How the reputation system functions in our focal marketplace

In our focal marketplace, when one party ends a contract both parties are prompted to give feedback.[2] 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 star 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 an 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 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 ratings now and in the past

The distribution of employer-on-worker feedback scores in our focal marketplace is highly right-skewed. Figure 1a depicts the histogram of public feedback scores from January 1, 2014 to May 11, 2016, for contracts worth more than $10.[3] 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 dashed line shows the cumulative

[2] We 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).

[3] We use this $10 restriction throughout the paper to remove mistaken, trial, and erroneous transactions.

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.

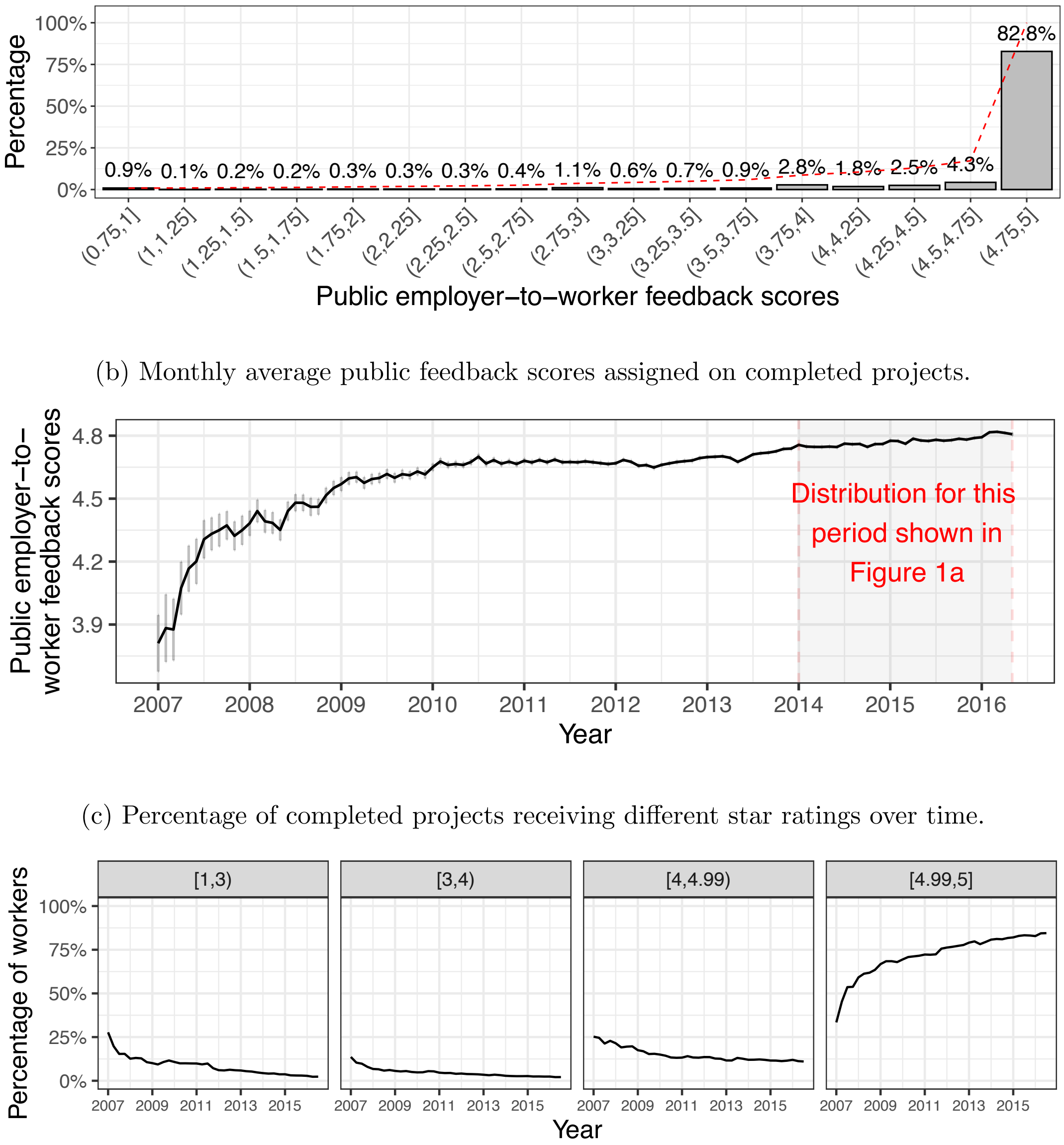
2.3 Evidence of increasing average scores from several online marketplaces

Our focal marketplace clearly shows an increase in average ratings over time, but a natural question is whether this kind of pattern is common in online markets. To answer this question, we collected average feedback score ratings from a number of online marketplaces. The average feedback scores for the various marketplaces are shown in Figure 2. For some marketplaces that are organized by geography, we also obtain city-specific data.

We observe an increase in average ratings over time that mirrors the pattern that we found in our focal market. While the precise slope and size of the increase differs by market (and by city), the general pattern of increase is clear.

The common pattern of increase in average feedback occurs despite the fact that the goods and services that are transacted in these marketplaces differ dramatically, and even though these platforms greatly differ in the marketplace mechanisms and matching technologies they employ. Panel (a) shows longitudinal data in a competing online labor market and ratings are assigned by employers to workers (freelancers). 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). The ratings are by guests (those who are renting properties) to hosts (those who are renting out properties). 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., 2021). The ratings are by users (renters of the

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



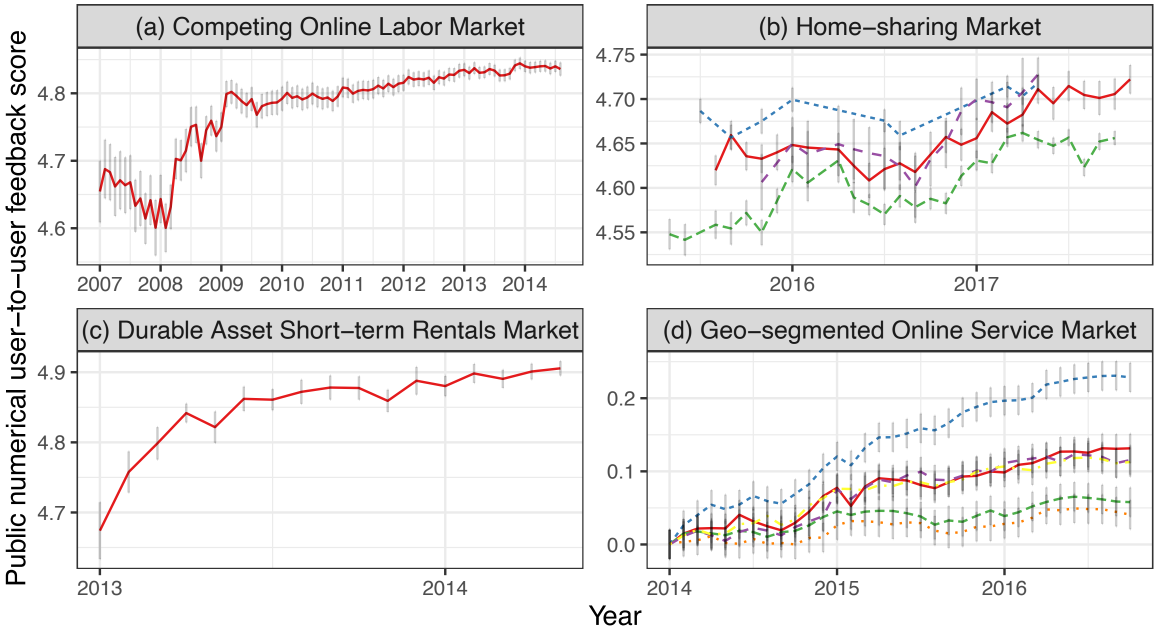
(a) Distribution of feedback scores for the period January 1, 2014 to May 11, 2016.

(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.

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.

asset) to users (providers of the asset) after the transaction has taken place. Panel (d) plots longitudinal ratings data from six major cities in the United States in a large online marketplace for services (Hall et al., 2021). The ratings are consumers of the service to providers of the service.

Despite the differences in the goods being transacted and the market mechanisms used, these marketplaces do share similarities. Transactions in these marketplaces are personal (peer-to-peer rather than person-to-firm). Furthermore, the same basic reputation system design is used across markets—ratings are given after the transaction has taken place and are consequential for the rated party, and all platforms use simultaneous reveal to prevent tit-for-tat rating behavior.

3 Descriptive evidence for reputation inflation

The increase in average feedback scores in a market could be explained by two broad—but not mutually exclusive—sets of reasons: (1) rater satisfaction has increased, and (2) reputation inflation, that is, raters are not any more satisfied but simply give higher feedback scores. If reason (2) is important, then it should leave some clues in the data; in this section, we examine some of these clues.

3.1 Some employers are not very satisfied and report strategically

If improvements in platform fundamentals have left raters very pleased, we might expect alternative measures of rater satisfaction to show similar increases, at least in direction. Of course, there is no immediate mapping from one measure of satisfaction to an other, but if a person gives "5 stars" in public but reports "it was not very good" in private, then one might suspect that the private measure is perhaps closer the rater's true feelings. The expressions "don't shoot the messenger" or "I dare you to say that to my face" are suggestive of why we might get more candor in private than in public.

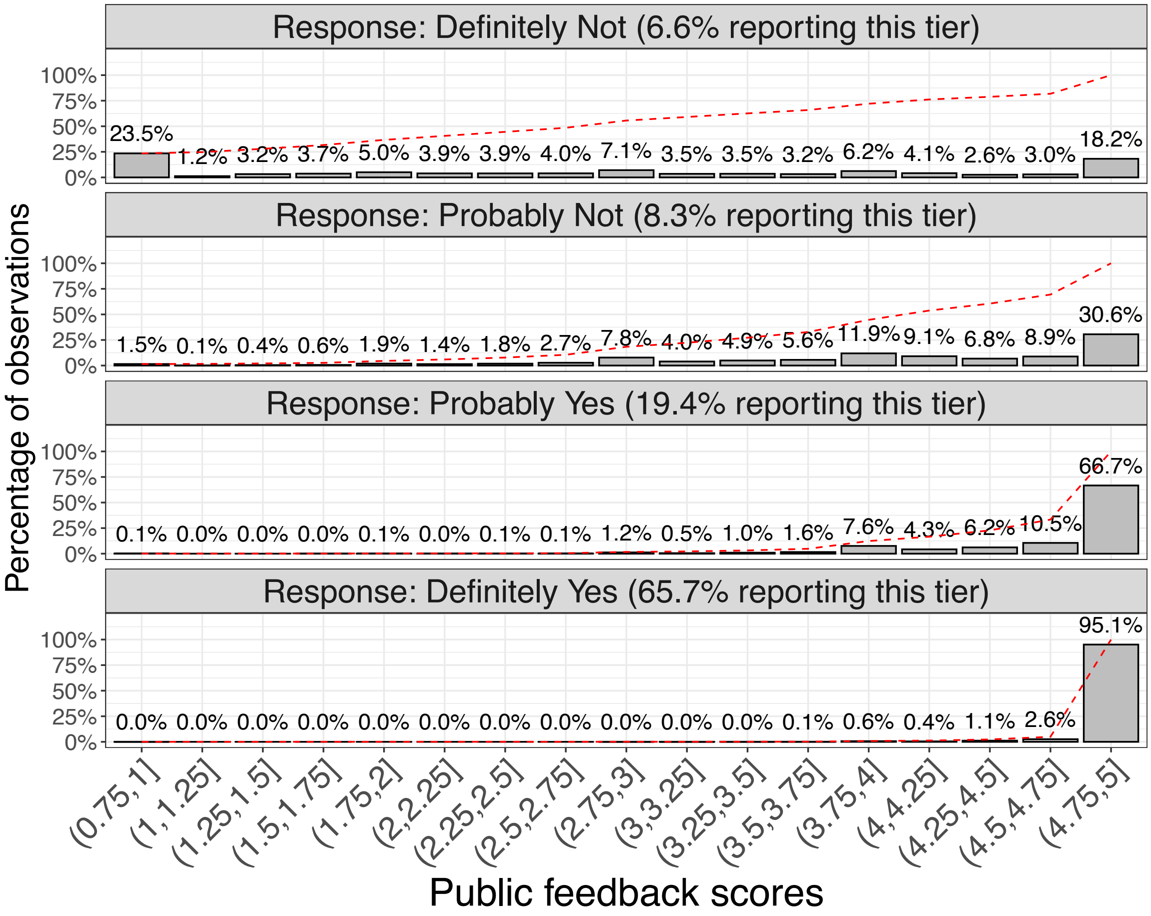
Toward that end of receiving more candid evaluations, the platform running our focal marketplace conducted an intervention that elicited an additional "private" feedback measure of satisfaction. This feedback measure was private in the sense that the platform let the employers know that private feedback would not be shared with the workers or with other employers, and that it would only be collected by the platform for internal evaluation purposes, such as to determine whether their recommendation systems needed to be improved. As with public feedback, this private feedback was elicited at the completion of a contract and was asked in addition to public feedback.

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?" The private feedback question was simply appended to the end of the public feedback form. Employers assigned both public and private feedback for the same contract.

Figure 3 shows the distribution of public feedback, conditioned on the private feedback. The percentage of employers giving that feedback score is shown in parentheses in each panel's label. Although the most common response to the question of "would you hire this freelancer again" 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 assign lower public feedback scores,

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?"



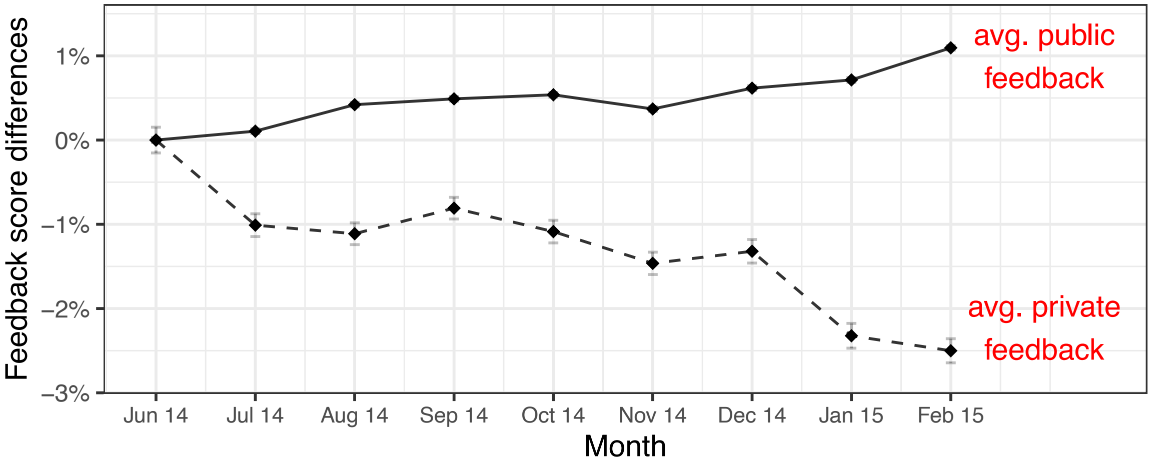
Notes: This figure plots the distribution of public feedback scores, computed separately for every set of users that gave the same answer to the private feedback question. The red dotted line plots the cumulative distribution function.

but many still give perfect public feedback scores. Among employers who selected the "Definitely No" answer to the private feedback question, 29.1% assigned a 1-star rating publicly. 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: none of these employers assigned less than 3.75 stars, and more than 95% of the observations falling into the highest bin.

3.2 Average private feedback sentiment decreased while public feedback sentiment increased for the same transactions

If private feedback scores were measuring rater satisfaction, then we would expect these two measures to co-vary over time: fundamental improvements that made raters happier should "show up" both in public and private ratings. In contrast, if one measure of satisfaction was inflating but the other was not, we could see a divergence.

Figure 4: Numerical public feedback score and private feedback score.



Notes: This figure 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. A 95% confidence interval is shown for each mean.

In Figure 4, we see a divergence between public and private scores, with public scores rising while private scores were falling. The figure reports the average monthly feedback over time, for the numerical public and private feedback (when the private feedback scale was numeric and 1 to 10). The x-axis covers the period when both were collected. To make the two scores comparable, we normalize them by their respective mean in the first period when they were both collected, i.e., $(\bar{s}_t - \bar{s}_0)/\bar{s}_0$, where $\bar{s}_t$ is the average feedback in month t. Public feedback scores exhibit a small increase during the period of interest (as we saw in Figure 1b), whereas 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 note that the average feedback scores shown in Figure 4 are being assigned by *the same employers on the same contracts.* The decreasing private feedback scores would seemingly suggest a decline in rater satisfaction, and yet public feedback scores increased. This divergence in trends suggests that the public feedback scores were increasing at least in part due to reputation inflation, assuming the private feedback score was not *deflating*.

An alternative explanation for the divergence is that the elicitation of private feedback somehow affected how employers assigned other types of feedback. For example, suppose employers who had negative experiences assigned workers bad private feedback scores instead of a bad public feedback scores, perhaps to "blow off steam." However, we view this as unlikely, as the private feedback was elicited simply by appending one additional question at the end of the feedback screen (see Appendix A.1). Even if employers read further down the page and considered both feedback decisions jointly, we would expect to see either a discontinuity in public feedback score averages, or a change in the rate of their increase, when private feedback was first elicited. We see no such pattern in the public feedback scores (see Figure 1b) or in the sentiment expressed in written employer feedback (which we will show in Section 4). We provide additional robustness checks in Appendix A.1, that rule out other conjectures that could rationalize the divergent trends, such as the possibility that workers misunderstood and misused private ratings.

3.3 The same written sentences are associated with much higher ratings now than in the past

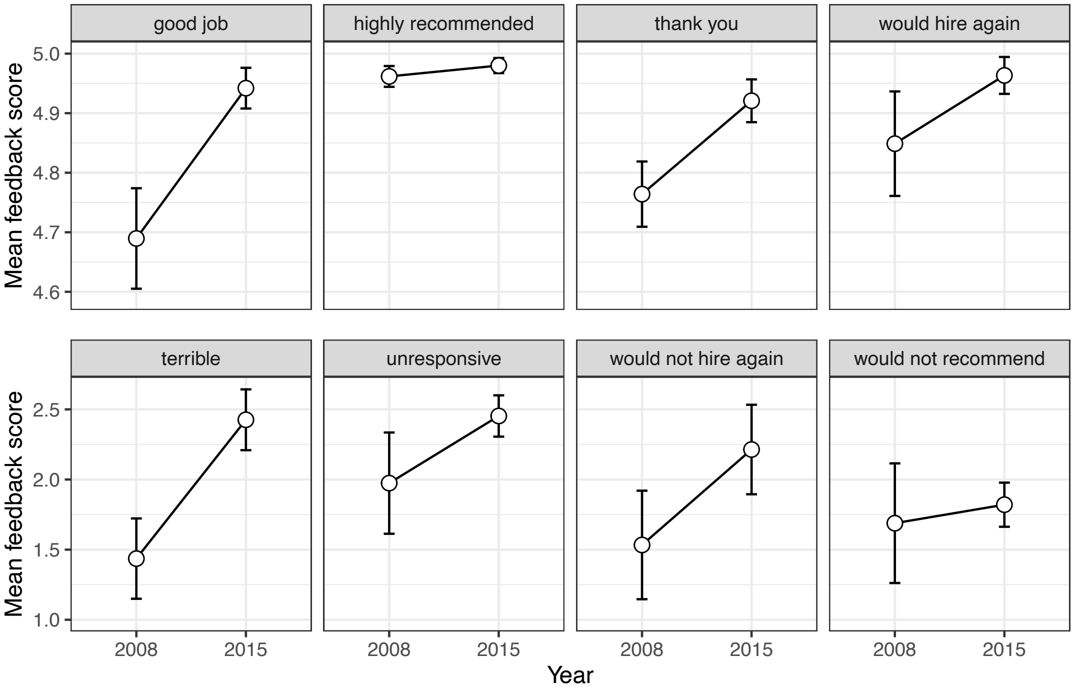
The telltale sign of reputation inflation is raters rating more positively without being more satisfied. An alternative measure of rater satisfaction can come from the written feedback employers leave after each transaction. If we think a distinct piece of written feedback—say "good job"—reflects an unchanging level of rater utility, then we can see if the numerical rating has changed over time for this phrase. To this end, we select written feedback from 2008 and 2015, and find all lexically identical sentences generated during these periods. We then compare average feedback by sentence across the two periods. Figure 5 shows the average numerical feedback scores for a set of commonly used short sentences across these two periods. We select sentences spanning both good and bad feedback, and which most frequently occurred in the corresponding written feedback in our data.

Figure 5 shows 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. This pattern is consistent with greater reputation inflation in the numerical feedback score.

4 Quantifying the contribution of reputation inflation

The private feedback and written text comparisons in Section 3 suggest an approach to quantifying reputation inflation: find some alternative measure of satisfaction and compare that to the primary measure suspected of inflating. We formalize this approach and show how using alternative feedback measures circumvents the problem of estimating latent utilities

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.

from observational data, and hence allows us to net out increases in average feedback scores due to improvements in marketplace fundamentals. We then use our method to quantify the importance of reputation inflation in our focal marketplace using written feedback.

4.1 Method for decomposing changes in average feedback scores

Let u denote the utility a rater obtains after some transaction.

Assumption 1. *In response obtaining utility u, the rater leaves primary feedback.*

$$s = \sigma(u) + \epsilon \tag{1}$$

where $\sigma(\cdot)$ is common among raters and monotonically increasing in the latent utility, i.e., $\sigma'(u) > 0$ for all u. Furthermore, $\mathbf{E}[\epsilon|u] = 0$ for all u, and so the expected score conditional upon the latent utility is

$$\mathbf{E}[s|u] = \sigma(u). \tag{2}$$

.

In addition to this primary feedback, raters also leave alternative feedback on the same transaction.

Assumption 2. *The alternative feedback is such that after each transaction, $a = \alpha(u)$, where $\alpha'(u) > 0$ for all u.*

Assumption 1 describes a data-generating process that is arguably a pre-condition for a useful reputation system. If different subjective ratings could not be usefully aggregated to some common scale, it is unclear how the system could convey information. This data generating process seems particularly appropriate for market settings where ratings do not simply reflect variation in taste but rather depend on some notion of satisfaction, given what was paid for the good or service. Assumption 2 simply states that the alternative measure is also a good measure—that when the latent utility is higher, it does "show up" with a higher score for that measure. Importantly, note that the monotonicity assumption ensures the inverse function, $\alpha^{-1}(\cdot)$, exists.

Suppose that for a set of transactions we observe a primary feedback set, S, and an alternative feedback set, A, but we do not observe the latent utilities, U. These are all data about the same transactions, with each transaction characterized by a tuple (s, a, u).

Assumption 3. *We can approximate the conditional expectation function (CEF) with a learned function $\hat{s}(a)$ such that*

$$\hat{s}(a) = \mathbf{E}[s|a] + \eta \tag{3}$$

where $\mathbf{E}[\eta|a] = 0$ for all a.

Assumption 3 implies that for all u,

$$\mathbf{E}[\eta|u] = 0, \tag{4}$$

or that there is not systematic error at any given utility.[4] Assumption 3 is certainly untestable. However, approximating the CEF without systematic error is precisely what flexible, modern machine learning methods are trying to accomplish. And given that the supports for both s an a are typically unchanged from one period to the next and the amount of ratings data can be vast in online marketplaces, this kind of prediction exercise is likely to go well in practice.

Now suppose that at some later time, we observe a new set of primary and alternative feedback, S' and A', with unobserved utilities, U'.

[4] If this was not the case, then there would be some u' such that $\mathbf{E}[\eta|u'] = k$ and $k \neq 0$. But $\mathbf{E}[\eta|u'] = \mathbf{E}[\eta|\alpha^{-1}(u')] = k$, which contradicts Assumption 3.

Proposition 1. *The expected value of the CEF applied to the new data is unbiased estimate of the expected score, regardless of the change in underlying utilities.*

Proof. The expected primary feedback score with the new data is $\mathbf{E}[s|U'] = \mathbf{E}[\sigma(u')|U']$, by Equation 1. If we apply the learned $\hat{s}(\cdot)$ on the new alternative feedback, the expected value is

$$\begin{aligned}
&= \mathbf{E}[\hat{s}(a')|A'] \\
&= \mathbf{E}[\hat{s}(\alpha(u'))|U'] \quad \text{(by Assumption 2)} \\
&= \mathbf{E}[\sigma(\alpha^{-1}(\alpha(u'))) + \eta|U'] \quad \text{(by Assumption 3)} \\
&= \mathbf{E}[\sigma(u')|U'] \quad \text{(by Equation 4)} \\
&= \mathbf{E}[s|U'] \quad \text{(by Equation 2 and iterated expectations).}
\end{aligned}$$

□

Proposition 2. *The expected inflation in the primary feedback score is the difference in the average rating and the expected rating using the learned CEF.*

Proof. If there is inflation in the primary metric—where $s = \sigma(u) + \tau(u) + \epsilon$, where $\tau(u) > 0$, we can obtain the average inflation by

$$\mathbf{E}[s|U'] - \mathbf{E}[\hat{s}(a')] = \mathbf{E}[\tau(u)|U']. \tag{5}$$

□

It is straightforward to show that if the alternative feedback measure also inflates, then our method yields a lower bound estimate. Our decomposition is conceptually similar to estimating monetary inflation (Sidrauski, 1967; Friedman, 1977; Mishkin, 2000; Berentsen et al., 2011), with the assumption that a "basket-of-goods" offers the same utility regardless of when it is consumed (Diewert, 1998). The difference in our setting is we can account for changes in the "quality" of the goods by using the alternative measure—something that is typically not possible in the monetary inflation case. As an aside, quality differences are a large conceptual issue for measure monetary inflation.[5]

Example decomposition Consider a platform where workers with types $\theta \in \{H, M, L\}$ produce goods with utilities u_H, u_M, and u_L respectively, with $u_H > u_M > u_L$. Employers match with workers, and receive a good and its corresponding utility. Employers then leave

[5] 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).

primary feedback $\sigma(u)$, such that $\sigma(u_H) = 1$, $\sigma(u_M) = 0.5$, and $\sigma(u_L) = 0$, and alternative feedback that is written text. Suppose the text is such that the employer always says "good" when $u = u_H$, "ok" when $u = u_M$ and "bad" when $u = u_L$. Using transaction data, we can approximate the CEF via the learned function

$$\hat{s}(a) = \begin{cases} 1 & , \text{ text} = \text{"good"} \\ 0.5 & , \text{ text} = \text{"ok"} \\ 0.0 & , \text{ text} = \text{"bad"} \end{cases} \tag{6}$$

We presented $\alpha(\cdot)$ as being a single measure in Section 4.1, but it could be constructed as index from numerous inputs, such as whether some text contained a specific term. Notice that we do not have to observe the underlying utilities to learn the function $\hat{s}(a)$.

At some later point in time $t' > t$, assume that improvements in marketplace fundamentals have resulted in employers only obtaining utilities u_M and u_H with equal probabilities, say due to better matching or a compositional change in sellers. This could be a result of any of the factors we discussed above; for example, the platform could have improved its matching systems, enabling employers to never match with workers of type L, workers of type L could now be more experienced or exert higher effort, and hence produce goods with utilities u_M and u_H with equal probabilities, or all workers of type L could have exited the platform. We can neither observe the reasons behind the shift in platform fundamentals, nor the new distribution of employer utilities at time t'. However, we observe the primary and alternative feedback scores left by employers. Insofar as no employers have not shifted the rating standards, employers leave average primary feedback equal to 0.75, and equal fractions of "good" and "ok" alternative feedback.

The crucial observation for our approach is that we can use the alternative-to-primary mapping learned from the period t data to estimate what the primary feedback average "should have been" in period t'. In particular, because employers leave alternative feedback scores "good" and "ok" with equal frequency at time t', we can estimate that the primary feedback score should have been 0.75. If, instead, the observed primary feedback average at time t' is 0.9, then the remaining 0.15 increase cannot be explained by improvements in fundamentals. We then say that $0.15/0.4 = 37.5\%$ of the observed increase in the primary feedback between time t and t' is attributable to reputation inflation. This allows us to disentangle changes in fundamentals from reputation inflation in our data.

It is important to note that if the alternative feedback measure also inflates, this approach will yield a lower bound on the magnitude of reputation inflation. For example, assume that when employers experience utility u_M at time t', they inflate their alternative feedback, generating "good" and "ok" with equal probabilities. We would then estimate that the primary feedback average should have been equal to 0.875, and hence conservatively estimate that $0.025/0.4 = 6.25\%$ of the observed increase in the primary feedback is due to inflation.

4.2 Using written feedback to estimate the degree of reputation inflation

Using written feedback, 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 employer written feedback as the training set, and the associated numerical scores as the set of labels. One advantage of the written feedback is that, unlike private feedback, we have access to written feedback over the entire platform history. 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.

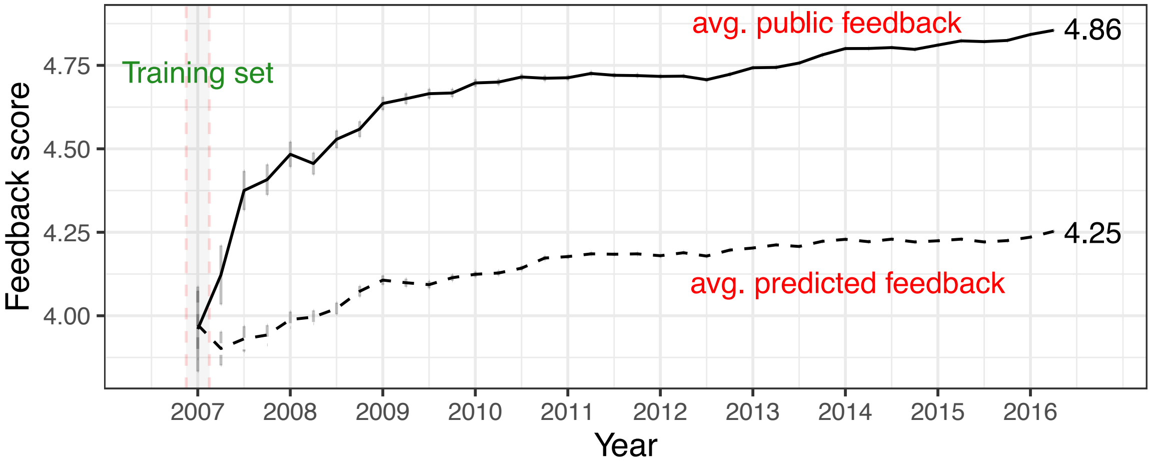
The average quarterly feedback scores over time, for both the numerical public feedback, and the feedback predicted from the written feedback, are plotted in Figure 6. As expected, the two scores match up during the training period. Going forward, both scores increase, but the predicted feedback score 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 the written sentiment and the 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.

One might be concerned that some kind of selection bias might be driving the divergence between public feedback scores and written text. We explore this possibility in Appendix A.2, finding no evidence of such a bias.

5 Discussion

Although we provided strong evidence that our alternative feedback measures—private feedback and written feedback—inflated at a slower rate (if at all) compared to public numerical feedback, we offered little as to why this might be the case. A related question is precisely why any measure inflates. In this section, we offer some thoughts on both questions, with an

Figure 6: Numerical public feedback score and predicted score from textual feedback



Notes: This 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 quarter from which training data was obtained for the predictive model.

eye towards future work that might be more definitive. We also discuss whether reputation inflation is likely to matter to the functioning of the reputation system.

5.1 Why might different measures of feedback inflate at different rates?

To quantify the importance of reputation inflation, we used two different alternative measures of rater satisfaction, private feedback and written feedback. We show that numerical public feedback inflates at a greater rate than these two measures, but a natural question is why the diference? We should note that our evidence cannot rule out that these other measures are also inflating—written feedback can certainly become inflated, with work that would have elicited a "good" now garnering a "great."[6] And of course, the private numerical feedback rating could also become inflated. However, it is important to reiterate that our method does not require there be no inflation in our alternative measure if the goal is to provide a lower bound.

A parsimonious explanation for why private scores are less prone to inflation is that these scores did not matter to the rated worker outcomes. As such, negative feedback by employers could not harm the rated worker. If employers want to avoid harm—either for altruistic reasons of because they expect some negative blow-back—they were relatively freer to give negative private feedback. Furthermore, because the rated worker did not know the

[6] One written feedback in our data reads: "This is the most impressive piece of coding in the history of software development!"

score, they could not complain.

Written feedback was of course public, but it might be less subject to less inflationary pressure than numerical feedback. First, it is harder for workers to complain about textual tone than it is to complain about a non-perfect star rating. Second, the platform does not aggregate written feedback or put it on a scale, making it harder to use than average numerical feedback for cross-worker comparisons by future employers; these comparisons are precisely what makes feedback consequential for workers. Third, the written feedback history is not presented in the worker's profile page in our focal marketplace, 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.

5.2 Causes of reputation inflation

Although we have strong evidence that reputation inflation exists, our data only hints at what the causes might be. The evidence suggests that rater unwillingness to give negative public feedback plays a role. Negative feedback is harmful to a rated party and raters might want to avoid that harm—either because they do not want to deal with retaliation (even just in the form of a complaint) or because they simply do not want to harm the rated party out of altruism. Recall that 28.4% of those employers who *privately* report that they would definitely not hire the same worker in the future, *publicly* assign them 4 or more stars out of 5. Because private feedback is anonymously given, workers cannot retaliate against employers following a bad private feedback score. At the same time, a bad public rating would be consequential in our setting, but a bad private rating would not.

While fear of retaliation or avoidance of harm—what we might think of as the cost of giving bad feedback—could explain a bias towards higher ratings, how does it explain the trends we observe? Although we do not model the process formally, it is easy to see how a kind of "ratchet effect" could happen in practice, with the cost of "bad" feedback rising over time. Suppose most raters want to rate "truthfully," that is, they want to "match" the percentile of their rating to the percentile of their subjective utility. I.e., if raters think their experience gave them the median level of utility, they would want to give the median feedback score; if they think they got the 25th percentile in utility, they want to give the 25th percentile score, and so on, even if this truthfulness can be harmful in the case of bad performance. But now suppose that some raters always just give the highest possible score, and avoid the costs of being truthful. These "always 5 stars" raters will shift the distribution of feedback scores, requiring even truthful raters to rate higher. This in turn will make any previous score fall in the distribution (e.g., 4 stars used to be the 80th percentile and now it's the 25th), effectively raising the cost of giving that score. Muchnik et al. (2013) design a large randomized experiment and find evidence supporting a similar mechanism.

5.3 Does reputation inflation matter?

Reputation systems exist to affect decisions in the marketplace and, indirectly, to create good incentives. A natural question is whether reputation inflation actually affects these system goals. It is not obvious that it would—for example, a certain amount of monetary inflation is desirable and creates no large loss in information so long as parties know to adjust. However, this is a misleading analogy. Unlike in monetary systems where there is no highest price, ratings systems are always on a top-censored scale: for the question "rate on a scale from 1 to X," the value of X must be pre-specified. This is why reputation inflation differs from monetary inflation: a sandwich that used to cost \$0.50 and may now cost \$12. However, this necessary increase could not happen if price was mechanically restricted to be below \$1. As such, reputation inflation leads to pooling of feedback scores in the highest feedback "bin." This pooling makes it difficult to distinguish "excellent" from simply "good," and with sufficient inflation the reputation system could become nearly binary, with the only possible signals being "terrible" and "not terrible." In our context, 85% of the users receive a perfect rating at the end of our sample (see Figure 1c), even though it is highly unlikely that 85% of transactions result in the exact same employer satisfaction.

Even if market participants and the platform is aware of the inflation, pooling is difficult or even impossible to correct statistically. With pooling, the strictly monotone relationship between rater satisfaction and scores is lost and presents the same problem as grade inflation (Babcock, 2010; Butcher et al., 2014). Furthermore, a strong rate of inflation—even if episodic and then contained—can cause individual reputations to vary based on when a feedback score was assigned, in turn undermining the usefulness of comparisons of feedback scores across different time periods. Aside from the effects on market participants, reputation inflation causes the platform itself to lose a yardstick for measuring its own performance.

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 a widespread problem.

For would-be marketplace designers, our paper illustrates a core market design problem. The diverge of public and private feedback scores in our data suggests that a possible mechanism driving reputation inflation is that raters incur a greater personal cost—or guilt—the greater the harm they impose on the rated worker. An interesting next step would be to elucidate the root causes of reputation inflation.

Whether there are effective market design responses to reputation inflation is an open question. Changes in the reputation system, such as adding a higher ceiling in the feedback scores, may temporarily mitigate—but do not solve—the problem.[7] Platforms could 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.[8]

Whether reputation systems less prone to inflation can be designed remains an open research question (Garg and Johari, 2020). Some of their problems seems to be manifestation of Campbell's law, which may be challenging to fully transcend (Campbell, 1979): "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."

[7] See also `https://www.youtube.com/watch?v=KOO5S4vxiOo`.

[8] For example, officer evaluation reports in the US Army limit senior raters to indicating only 50% or less of the officers they rate as "most qualified." However, it may be challenging to force a distribution in settings where buyers evaluate sellers as a "flow."

References

Aperjis, Christina and Ramesh Johari, "Optimal windows for aggregating ratings in electronic marketplaces," *Management Science*, 2010, *56* (5), 864–880.

Aral, Sinan, "The problem with online ratings," *MIT Sloan Management Review*, 2014, *55* (2), 47.

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.

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.

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.

Campbell, Donald T, "Assessing the impact of planned social change," *Evaluation and Program Planning*, 1979, *2* (1), 67–90.

Dellarocas, Chrysanthos, "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.

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*, 2020, *66* (9), 4152–4172.

_ , Srikanth Jagabathula, and Arun Sundararajan, "The limits of centralized pricing in online marketplaces and the value of user control," *Working paper*, 2021.

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.

Garg, Nikhil and Ramesh Johari, "Designing Informative Rating Systems: Evidence from an Online Labor Market," in "Proceedings of the 21st ACM Conference on Economics and Computation" EC '20 Association for Computing Machinery New York, NY, USA 2020, p. 71.

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.

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.

_ , John J Horton, and NBER Daniel T Knoepfle, "Pricing in Designed Markets: The Case of Ride-Sharing," *Working paper*, 2021.

Horton, John J, "Online labor markets," *Internet and network economics*, 2010, pp. 515–522.

_ and Prasanna Tambe, "Labor economists get their microscope: big data and labor market analysis," *Big data*, 2015, *3* (3), 130–137.

_ , 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.

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.

Levin, Jonathan D, "The Economics of Internet Markets," Technical Report, National Bureau of Economic Research 2011.

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.

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

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.

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.

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 More details on the alternative feedback method

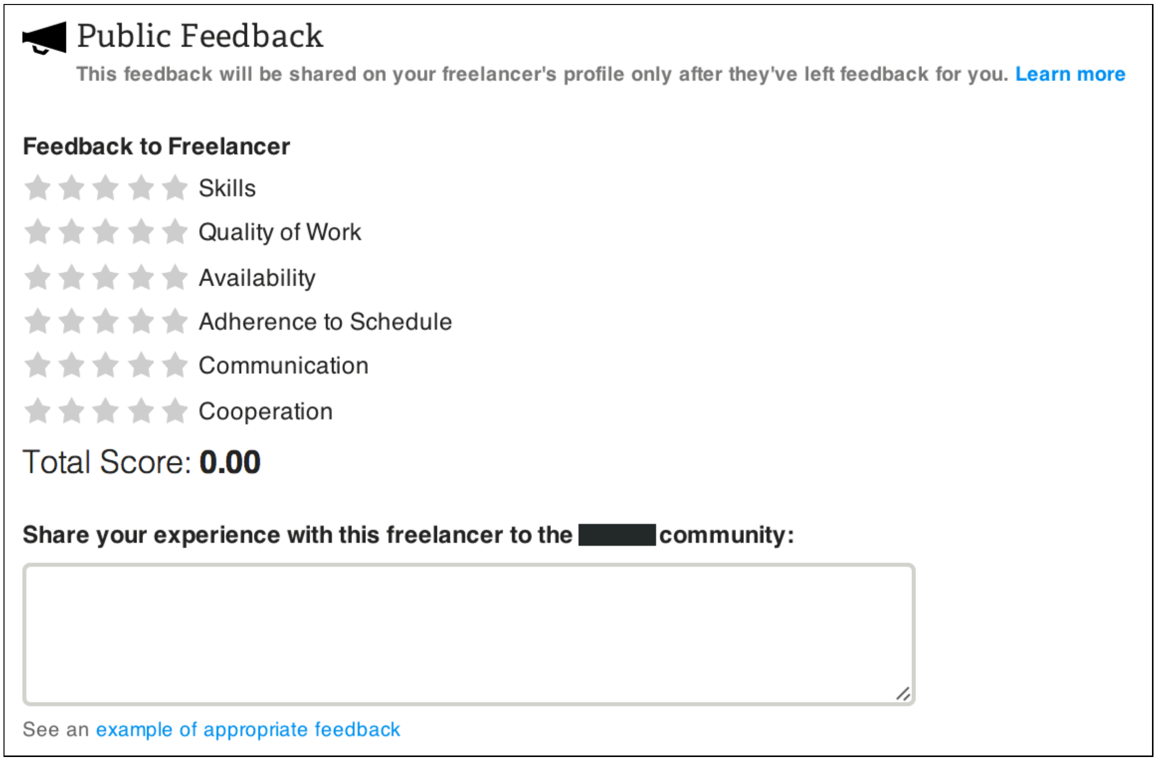
A.1 Robustness tests for private feedback

A.1.1 Details on private feedback elicitation

The status-quo employer-on-worker feedback is shown in Figure 7a. For private feedback solicitation, the interface shown in Figure 7b 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 7: 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

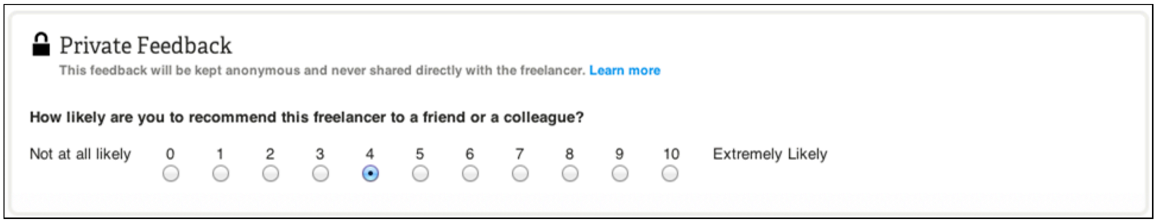
☆☆☆☆☆ Cooperation

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

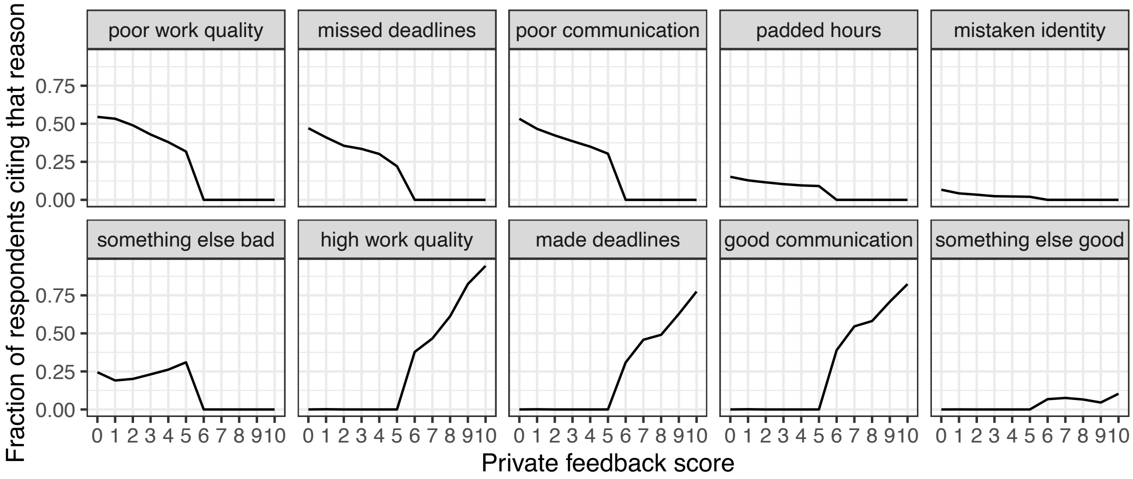
How likely are you to recommend this freelancer to a friend or a colleague?

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

A.1.2 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 8. 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 8: Fraction of users citing a given reason when giving private feedback, by score.

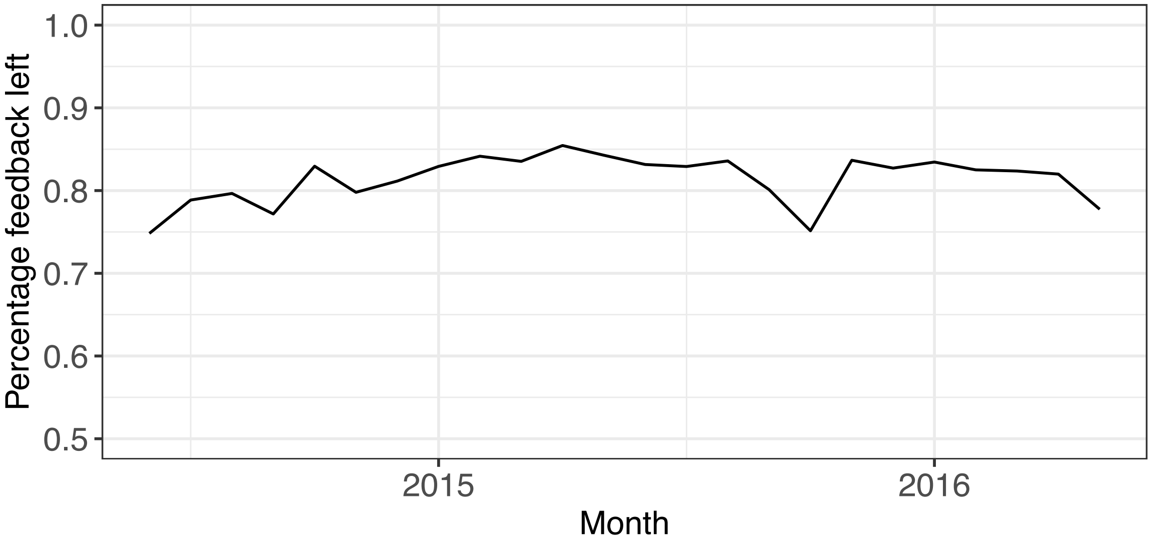


Notes: This figure plots the fraction of feedback reports that cited each reason as the basis of the feedback being positive or negative, against the private feedback score given. Across every case, we notice that employers that assigned more extreme feedback scores tend to cite reasons of the same sentiment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

A.1.3 Selection

A plausible concern is that the employers' decision to leave private feedback when they leave public feedback could change over time. Figure 9 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. The percentage of employers that chooses to leave private feedback is high—81.4% of employers decide to also assign private feedback.

Figure 9: 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.

A.2 Robustness tests for written feedback

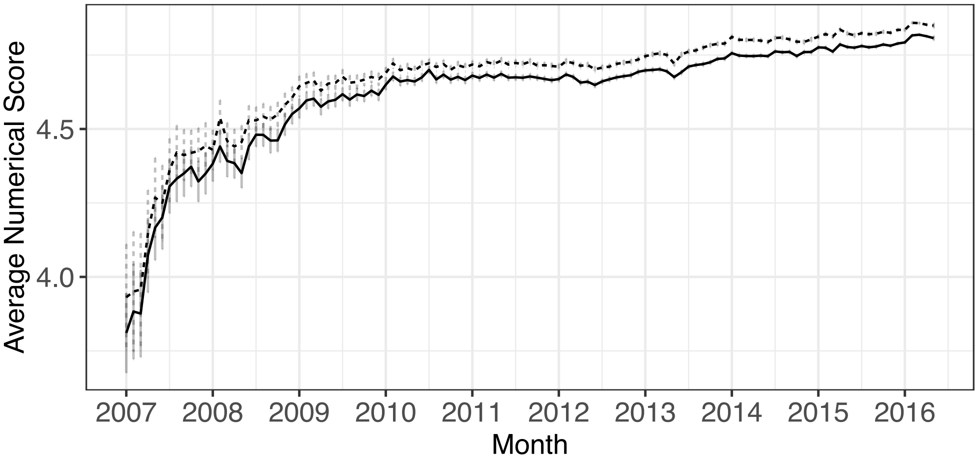
A.2.1 Selection issues

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 10 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.

Similarly to private feedback, a concern is that employers decision to leave written feedback when they leave public feedback could change over time. Figure 11 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

Figure 10: 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.

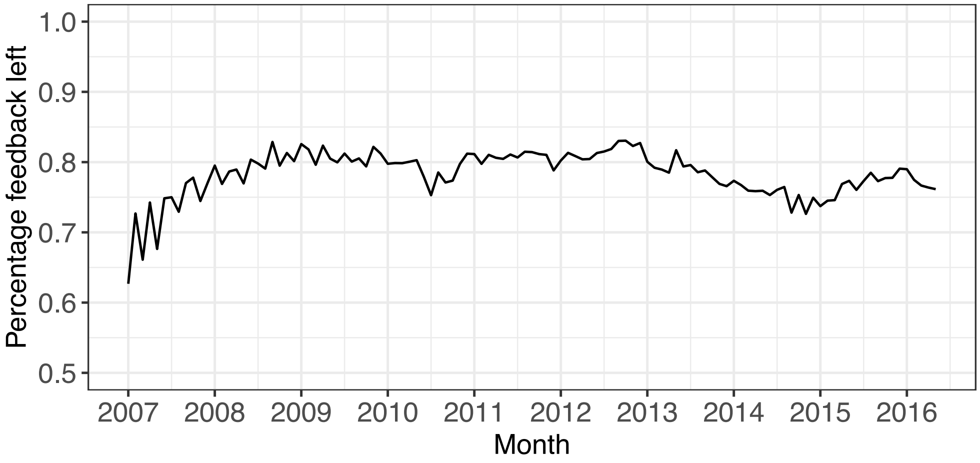
assign private feedback when they assign public feedback. The percentage of employers that chooses to leave written feedback is also high, with an average of 79.2% of employers deciding to also assign written feedback.

A.2.2 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 4.1, this issue can be thought of as changes in the conditional expectation function.

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 $\overline{r}_T = \sum_{\text{i left feedback in T}} r_i$. We then test whether, among 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

Figure 11: 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.

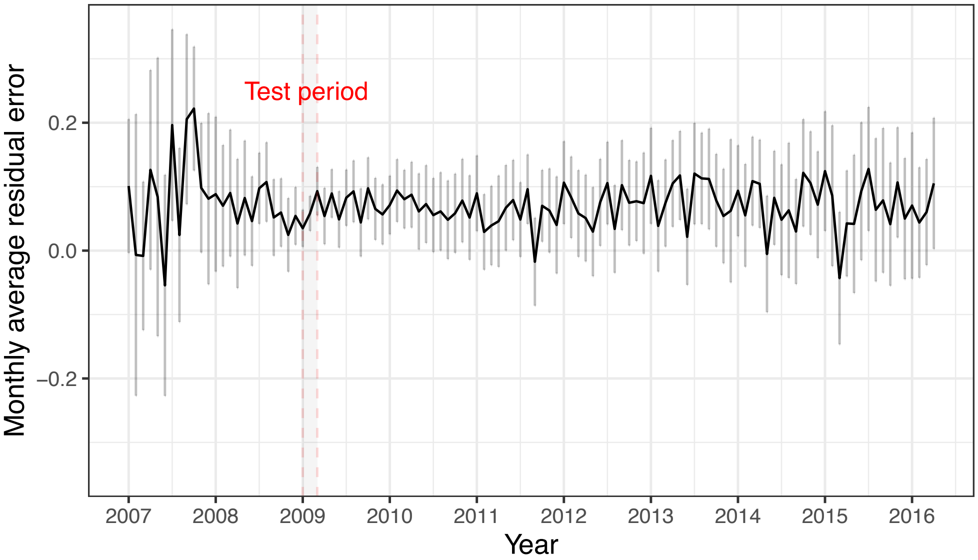
can do so by simply computing $\overline{r}_t = \sum_{\text{i left feedback in T and t}} r_i$, for any $t \neq T$. If for $t > T$ the quantities $\overline{r}_t$ show a systematic increase, then this composition shift in rater types may bias our estimates.

Figure 12 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 13. 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.

A.2.3 Alternative training periods

In the bottom panel of Figure 13 we perform the same empirical exercise as in Section 4.2, 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

Figure 12: Employer average residual error in for employers who left feedback during January and February 2009.



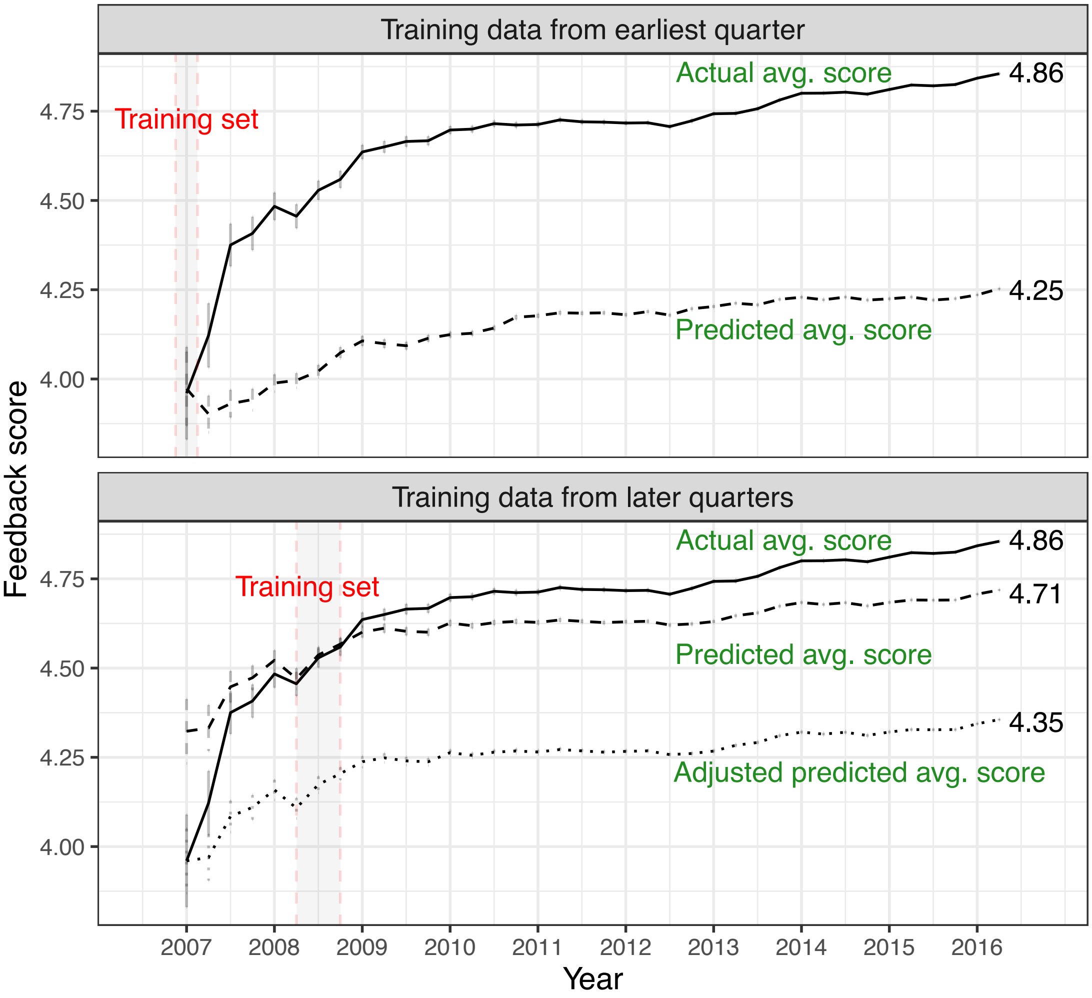
Notes: This figure plots the employer average residual error over time for the set of employers who left feedback during the period indicated by the shaded area. The average residual errors are computed for every month, and a 95% interval is depicted for every point estimate.

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.

A.2.4 Predictive algorithm performance

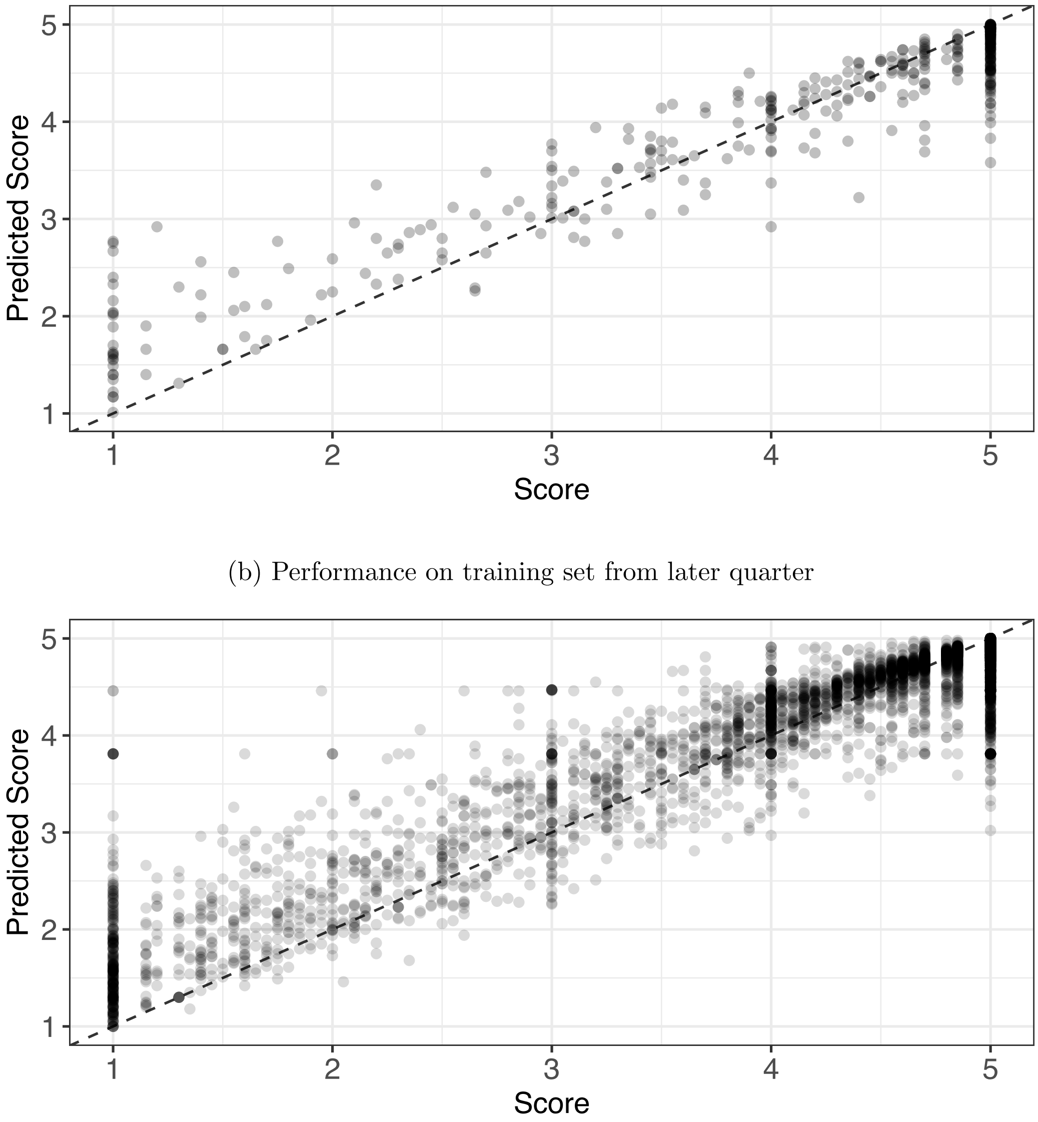
We present more details about the performance of the algorithms used to extract the written feedback sentiment in Section 4.2. Figure 14a plots the scatterplot of numerical scores versus predicted scores from written feedback for the algorithm trained on data coming from the earliest quarter. Figure 14b 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 13: 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.

Figure 14: 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.