Rater Altruism Reduces the Efficiency of Reputation Systems[*]

Apostolos Filippas
Fordham

John J. Horton
MIT & NBER

November 3, 2023

Abstract

Feedback in reputation systems often becomes more positive over time, even without improvements in rater satisfaction. This "reputation inflation" undermines the efficiency of reputation systems, both online and offline. To examine the mechanisms behind this phenomenon, we develop a model where raters increasingly inflate the feedback they give because (i) it is costly for raters to give bad feedback to ratees, and (ii) what constitutes bad feedback is endogenous—it depends on what feedback other raters give. Our model explains prevalent temporal patterns within reputation systems, as well as variations between reputation systems in different contexts. The costs that raters face could stem from raters fearing that ratees will retaliate, but also because raters altruistically worry that a bad rating might harm the ratees. Using a unique quasi-experiment where the costs of receiving bad feedback increased for ratees but the identity of the raters remained anonymous, we show that the rater altruism alone may cause reputation inflation. Our results elucidate the reasons why reputation systems are prone to reputation inflation in personal settings, such as peer-to-peer platforms. We contribute to the large literature on reputation system design with, to our knowledge, the first empirical evidence form the field to show that rater altruism is enough to erode the effectiveness of reputation systems. Our findings are directly relevant to market designers of reputation and rating systems in online or offline marketplaces.

[*]The latest version of this paper can be found at https://www.apostolos-filippas.com.

1 Introduction

The practical value of reputation systems decreases considerably if raters assign ever higher feedback irrespective of their satisfaction. Several reputation systems, both online and offline, suffer from this kind of inflation. For example, in US colleges, the median grade is often the highest possible grade (Rojstaczer and Healy, 2012); in online marketplaces, anything less than a perfect score typically constitutes bad feedback (Filippas, Horton and Golden, 2022).

This paper focuses on identifying the cause of reputation inflation. Our starting point is to examine the economics of feedback systems. First, raters typically incur higher costs for giving "bad" feedback than for giving "good" feedback. This cost divergence could explain a kind of "positivity" bias—more good feedback and less bad feedback—but it cannot explain the inflationary process where bad feedback becomes increasingly rare over time. Second, what constitutes "bad" feedback—feedback that causes worse market outcomes for ratees who receive it—is not immutable; rather, it is context-specific, and it depends on the inferences others draw in that particular reputation system and market. For example, two stars in the Uber app is a disaster; but two stars in the Michelin guide is a triumph.

We formalize these observations in a model of a marketplace with a reputation system. Our model has three basic elements. First, ratees incur a cost from receiving bad feedback, in the form of worse future market outcomes. Second, what constitutes bad feedback is endogenous, depending on the on the current distribution of feedback. Third, raters have a preference for being truthful, but they also incur "reflected" costs for giving bad feedback. These reflected costs are increasing in the the ratees' costs from receiving bad feedback. We show that there exists an equilibrium in which reputations are universally inflated.

The key driver of the universal inflation prediction is the possibility of reflected costs. In any actual reputation system, the two main sources of reflected costs are (1) ratee retaliation and (2) rater altruism. Raters may incur reflected costs for giving bad feedback because they worry that ratees will retaliate against them. For example, angry ratees may send complaints or withhold future collaboration, and would-be trading partners may avoid someone who has a "strict rater" reputation. However, raters may also incur reflected costs because they are altruistic, that is, raters do not want to harm the ratees' future prospects in the market. For example, a ride-sharing passenger might not be pleased with the route her driver took or dislike her driver's music taste, but she also does not wish to ruin the driver's livelihood with a bad rating.

The two possible cost sources—retaliation and altruistic concerns—are empirically hard to pick apart, but disentangling them is of great practical importance. If retaliation is the sole driver of inflation, then a reputation system that effectively hides the identity of the rater should be sufficient to prevent reputation inflation. But if rater altruism matters, anonymization is not enough to stymie reputation inflation. We explore this distinction using a unique quasi-experiment which allows us to disentangle ratee retaliation from rater altruism.

Our empirical context is an online labor market which introduced a new "private" feedback system in addition to its existing public feedback system. Importantly, the way the private feedback system was released to the marketplace changed the raters' costs of giving bad feedback at different times. At first, the private feedback scores were neither revealed nor used by the platform in any way or form. As such, raters incurred no reflected cost when giving a bad private feedback score. However, after nine months, the platform made private feedback consequential by revealing it publicly. Crucially, the platform revealed the private feedback in a *batched aggregate* manner: a ratee's private feedback score would only be updated each time she received five new private feedback scores from five different raters, and the identities of the raters would not be revealed. This batching provided quasi-anonymity to raters and hence it precluded retaliation as a reflected cost. However, the batching left open only the possibility of altruistic reflected costs.

We show that after the revelation of the private scores, private feedback scores began inflating immediately. The totality of evidence suggests that the raters' altruistic concern was the cause for this inflation. We rule out several other competing hypotheses that could also explain the sudden increase in private feedback scores, such as improved rater satisfaction and composition shifts. To achieve this, we exploit the fact that both a public and a private reputation system coexisted at the same time.

A simple interpretation of our results is that, on average, raters do not want to harm ratees. Although altruism is certainly laudable, the potential for inflicting harm is what gives a reputation system much of its "bite." Without this bite, the usefulness of many modern reputation or rating systems declines.

In terms of the generalizability of our findings, one seeming limitation is that some rating systems are not excessively positive or prone to inflation. For example, product reviews on Amazon, movie reviews on RottenTomatoes, and restaurants reviews on Yelp can be quite negative generally, and do not seem to have become notably inflated. Our theoretical model and our empirical evidence offer a parsimonious explanation for this difference. In highly personal settings where the rated party is an individual, such as in peer-to-peer platforms (Einav, Farronato and Levin, 2016; Filippas, Horton and Zeckhauser, 2020), altruism creates a reflected cost and hence we have inflation; but in impersonal settings where the rated party is a firm or cultural product, altruism concerns are minimal and so inflation does not occur.

A large body of research has examined reputation systems, both in online and offline markets (Bolton and Ockenfels, 2000; Li and Hitt, 2008; Tadelis, 2016; Chevalier, Dover and Mayzlin, 2018; Gutt, Neumann, Zimmermann, Kundisch and Chen, 2019; Fradkin, Grewal and Holtz, 2021; Kokkodis, 2021; Yang, Zheng and Mookerjee, 2021). We believe our work contributes to this area by offering a compelling theoretical analysis on the causes of reputation inflation, and empirical evidence supporting the view that rater altruism is a root cause for this phenomenon. Our work contributes to this literature by showing that in personal settings, designing reputation systems that are not prone to inflation is challenging.

The rest of the paper is organized as follows. Section 2 presents a model of reputation inflation, and derives general results about its equilibrium properties and implications. Section 3 examines the quasi-experimental revelation of private feedback information. Section 4 discusses the implications of our findings. Section 5 concludes.

2 A model of reputation inflation

We develop a simple model of reputation inflation, where raters decide whether to be candid and assign truthful feedback, or to lie and assign inflated feedback. The rater's decision depends on the ratee's cost of receiving truthful feedback, which in turn depends on the degree of reputation inflation in the market—an equilibrium object.

2.1 Model primitives

Consider an online labor market composed of workers and employers. Employers match with workers randomly. After an employer-worker match forms, the following sequence of events takes place: (1) the employer observes the workers's reputation, and pays the worker wage w, (2) the worker produces output y, and (3) the employer receives the output, and assigns feedback s to the worker.

Workers are endowed with quality $q \in \{q_L, q_H\}$, with $q_H > q_L$, indicating high and low quality. Employers do not observe workers' qualities, but the fraction of high-quality workers is publicly known—we assume that this fraction is equal to 1/2 to simplify our calculations. Workers produce output $y \in \{\mathrm{G}, \mathrm{B}, \mathrm{T}\}$, indicating good, borderline, and terrible output. The production function of high-quality workers is $\Pr(y = \mathrm{G}|q = q_H) = q_H$, and $\Pr(y = \mathrm{B}|q = q_H) = 1 - q_H$. The production function of low-quality workers is $\Pr(y = \mathrm{G}|q = q_L) = q_L$, $\Pr(y = \mathrm{B}|q = q_L) = 1 - q_L - \alpha$, and $\Pr(y = \mathrm{T}|q = q_L) = \alpha$. Thus, any worker may produce good or borderline outputs, but only low-quality workers may produce terrible outputs. The outputs are strictly ordered by their values to employers, $v_H > v_B > v_T$, and hence the value distributions satisfy the strict monotone likelihood ratio property. The expected value of hiring a high-quality worker is u_H, and the expected value of hiring a low-quality worker is u_L, with $u_H > u_L$.

Employers assign feedback score $s \in \{0, 1\}$ to the worker after each transaction. This score is intended to indicate whether they received good output ($s = 1$) or not ($s = 0$), but what score they actually report is a choice up to each employer. In the case where employers always report feedback truthfully, the average feedback score will be $(q_H + q_L)/2$. Every time a new match forms, the employer observes the worker's most recent feedback, forms a belief about the worker's quality, and conditions the wage she pays to the worker upon that belief. Workers and employers are price-takers, and hence each worker is paid her expected marginal product,

$$w_s = \Pr\left(q = q_H|s\right) u_H + \Pr\left(q = q_L|s\right) u_L.$$

In words, $w_{s=1}$ is the wage associated with good feedback, and $w_{s=0}$ is the wage associated with bad feedback. Note that $u_H > w_{s=1} > w_{s=0} > u_L$, that is, high-quality workers are always paid less than their value, and low-quality workers are paid more than their value.

2.2 Reflected costs and feedback assignment

In deciding whether to inflate their feedback scores, employers take two factors into account. First, employers obtain a truth-telling benefit $b > 0$ when they report truthful feedback.[1] Second, employers incur a "reflected" cost $c_i \Delta w$ when they report truthful feedback, where

$$\Delta w = \Pr(q = q_H|y)(u_H - w_s), \tag{1}$$

and c_i is drawn from a publicly known distribution $F : [\underline{c}, \bar{c}] \to [0,1]$. The term Δw is the probability that the employer assigns feedback to a high-quality worker times the cost of that feedback to a high-quality worker, and the term c_i is an employer-specific reflection coefficient. Reflected costs come from the chance that the feedback is harming an actually good workers, times the damage this does to them compared to what they actually produce in expectation. Together, these two terms capture the fact that employers may differ in their propensities to inflate their feedback, and that assigning truthful feedback becomes costlier for employers as the workers' costs of receiving that feedback grow.

The "reflected" costs of assigning truthful feedback may include the employer's aversion to harming the rated worker's future prospects (altruism), as well as the costs of workers complaining, withholding future cooperation, and other workers being unwilling to work for employers with a "strict rater" reputation (retaliation). Furthermore, note that employers only consider the cost of their ratings to high-quality workers, because $w_s > u_L$ for all s.

Putting these factors together, employers report feedback truthfully if

$$b \geq c_i \Delta w. \tag{2}$$

Note that if employers do not care about costs—$c_i = 0$ for all employers—then there is no inflation.

It is worth examining more closely the structure that our assumptions impose on the feedback assignment process. Employers who receive a terrible output always assign bad

[1]The benefit parameter models platform-specific benefits such as awards for accurate reviews, as well as the rater's intrinsic motivation to be honest (Abeler, Nosenzo and Raymond, 2019).

feedback truthfully, because high-quality workers never produce terrible output, and hence $\Pr(q = q_H | y = \mathrm{T}) = 0$.[2] Some employers who receive good output would like to inflate their feedback because even at $s = 1$, a high-quality worker gets paid less than their marginal product, but the rater's hands are tied because $s = 1$ is the feedback upper bound (this type of "top-censoring" is a common characteristic of reputation systems). As such, employers always report good feedback truthfully upon receiving a good output. Employers who receive a borderline output may choose to lie, and inflate their feedback.

2.3 Equilibrium

In light of the possibility for inflated feedback scores, future employers will condition their wages on the likelihood that the observed feedback is truthful. Let p denote the fraction of employers that assign inflated feedback, and assume that p is common knowledge. Employers who observe bad feedback infer that

$$\begin{aligned}\Pr(q = q_H | s = 0; p) &= \frac{\Pr(s = 0 | q = q_H; p)\Pr(q = q_H)}{\Pr(s = 0; p)} \\ &= \frac{(1 - q_H)(1 - p)}{(1 - q_H)(1 - p) + (1 - q_L - \alpha)(1 - p) + \alpha},\end{aligned} \tag{3}$$

employers who observe good feedback infer that

$$\begin{aligned}\Pr(q = q_H | s = 1; p) &= \frac{\Pr(s = 1 | q = q_H; p)\Pr(q = q_H)}{\Pr(s = 1; p)} \\ &= \frac{q_H + (1 - q_H)p}{\big(q_H + (1 - q_H)p\big) + \big(q_L + (1 - q_L - \alpha)p\big)},\end{aligned} \tag{4}$$

and workers are paid wage

$$w_s(p) = \Pr\big(q = q_H | s; p\big)\, u_H + \Pr\big(q = q_L | s; p\big)\, u_L. \tag{5}$$

It is straightforward to show that $w_{s=0}(p)$ is concave decreasing in p. As feedback scores become more inflated, bad feedback becomes costlier for workers because it is more likely assigned following a terrible output—which only low-quality workers may produce. In contrast, $w_{s=1}(p)$ is convex decreasing in p if $\alpha < 1 - \frac{q_L}{q_H}$, and concave increasing in p otherwise. The role of α is determining the shape of $w_{s=1}(p)$ can be seen by considering an example. Consider the extreme case where low-quality workers produce terrible outcomes, but they never produce borderline outcomes ($\alpha = 1 - q_L$): they only produce good or terrible output. In this case, only high-quality workers receive inflated feedback when they produce borderline output, and

[2] While not explicitly modeled and treated as random, one might imagine that terrible output might cause the rater to want to punish the ratee for a kind of defection, (who, after all, was paid); this kind of equity/reciprocity type concern has a strong empirical basis (Bolton and Ockenfels, 2000).

hence good feedback correlates more strongly with the worker being a high quality worker.

Using Equation 1 but making it depend on p, let $\Delta w(p) = \Pr(q = q_H | y = \mathrm{B})(u_H - w_{s=0}(p))$. The equilibrium feedback inflation p^* is then found by solving the equation

$$p^* = 1 - F\left(b/\Delta w(p^*)\right). \tag{6}$$

An equilibrium exists for any continuous distribution function, but is not unique in general. The two extreme cases where

$$p^* = \begin{cases} 0, & \text{if } \; b > \bar{c}\Delta w(1) \\ 1, & \text{if } \; b < \underline{c}\Delta w(0) \end{cases},$$

correspond to corner solutions indicating the all-truthful and all-lying equilibria. If the benefit to assigning truthful feedback is higher than the cost for every employer (including with the highest reflection coefficient, $\bar{c}$), even if all employers were inflating, then all employers tell the truth and $p^* = 0$. Similarly, if the benefit of telling the truth is less than the cost for even the least cost-sensitive client ($\underline{c}$), then all employers inflate, and $p^* = 1$.

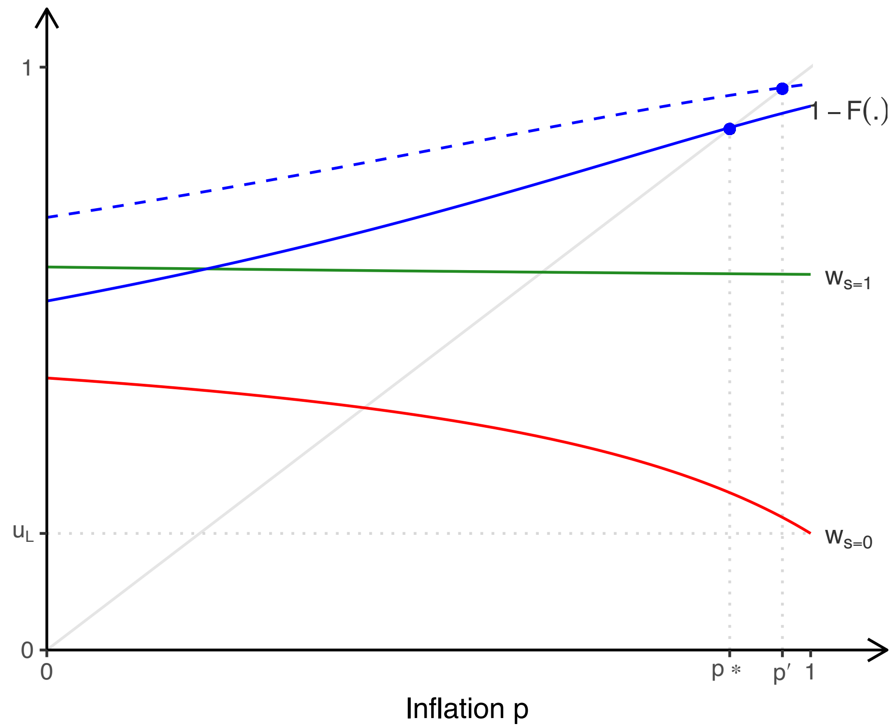
In most online marketplaces the benefit b could be small, and sometimes even zero, compared to say being a movie critic or a Yelp reviewer whose *own* reputation could suffer from giving untruthful reviews; by comparison, pulling punches and shading up the ratings for a Lyft driver is not going to cause anyone to question the rater's judgment and taste. At the same time, reflected costs can be substantial if a bad review is very harmful to the ratee. As such, to the extent that we think of employers as both strategic and narrowly self-interested, the all-lying equilibrium is the more likely outcome.

2.4 Graphical illustration

To illustrate the equilibrium of our model, Figure 1 depicts various quantities in our model as a function of the inflation p. The solid lines depict the case of $q_H = 0.8$, $q_L = 0.6$, $\alpha = 0.2$, $u_H = 1$, $u_L = 0.2$, $b = 0.25$, and $F \sim N(1, 0.25)$. The green solid line depicts the wage returns to good feedback, $w_{s=1}(p)$, and is convex decreasing in the degree of reputation inflation because $\alpha < 1 - \frac{q_L}{q_H}$. The red solid line depicts the wage returns to bad feedback, $w_{s=1}(p)$; it is concave decreasing in the degree of reputation inflation, and attains its minimum, u_L, at full inflation. With this parameterization, as inflation increases, there is a lower return to both good and bad feedback. The blue solid line depicts the fraction of employers who choose to inflate their feedback, $1 - F\left(b/\Delta w(p)\right)$. The unique equilibrium p^* is the value for which it crosses the 45-degree line (see Equation 6).

To see how a change in the reflection coefficients changes the market quantities, we increase the mean of the distribution F by 10%. The blue dashed line depicts the new fraction of employers who choose to inflate their feedback at each level of inflation p. The new equilibrium inflation, p', is about 7.8% higher than p^*.

Figure 1: Equilibrium example showing the relationship between wages from good and bad feedback for ratees, conditional upon the degree of reputation inflation by raters



Notes: This figure plots quantities of the model as a function of reputation inflation. The solid lines depict the case of $q_H = 0.8$, $q_L = 0.6$, $\alpha = 0.2$, $u_H = 1$, $u_L = 0.2$, $b = 0.25$, and $F \sim N(1, 0.25)$. The solid green line depicts the wage returns to good feedback, $w_{s=1}(p)$, and the solid red line depicts the wage returns to bad feedback, $w_{s=1}(p)$. The solid blue line depicts the fraction of employers who choose to inflate their feedback, $1 - F\left(b/\Delta w(p)\right)$. The equilibrium inflation p^* is then the point at which the solid blue line intersects the 45-degree line. The dashed blue line depicts the fraction of employers who choose to inflate their feedback when $F \sim N(1.1, 0.25)$, and with all other parameters fixed. In this case, the equilibrium inflation p' increases.

3 A quasi-experiment making feedback consequential

In the model of Section 2, the reflected costs that raters incur for giving bad feedback cause reputation inflation. A prediction of the model is that shifts in the cost borne by ratees should in turn shift the equilibrium fraction of raters inflating their feedback, and hence change the degree of reputation inflation in equilibrium. In this section, we test this prediction with a quasi-experiment caused by a policy change that increased the ratees' costs of receiving bad feedback, but did not allow the ratees to retaliate. Crucially, because the identity of the rater was kept anonymous there could be no rater retaliation, and hence the only possible source of reflected costs could be raters' altruistic concerns.

3.1 Empirical context

Our empirical context is a large online labor market (Horton, 2010; Agrawal, Horton, Lacetera and Lyons, 2015; Horton, Kerr and Stanton, 2017). In online labor markets, employers hire workers to perform tasks that can be done remotely, such as computer programming, graphic design, data entry, research, and writing. Each market differs in its scope and focus, but platforms commonly provide ancillary services that include maintaining job listings, hosting user profile pages, arbitrating disputes, certifying worker skills, and maintaining feedback systems (Benson, Sojourner and Umyarov, 2019; Filippas et al., 2020, 2022).

Historically, the platform had one kind of feedback, which we call "public" feedback. For this kind of feedback, the platform asks trading partners to assign each other feedback when their contract ends. Both parties have a common 14-day period in which to leave feedback. If both parties leave feedback before the deadline, 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. Neither party can assign feedback or change their assigned feedback past the deadline, and neither party learns the feedback it received before assigning feedback to the other party. As such, direct "tit-for-tat" conditioning is not possible (Bolton, Greiner and Ockenfels, 2013). Leaving feedback is strongly encouraged and exceeds 80%, but not compulsory. We focus on employer-to-worker feedback in what follows.

The employer public feedback has two parts. Employers assign public numerical feedback by rating the worker on a 1-5 star scale across several dimensions, which are then aggregated according to publicly known weights. Employer assign public written feedback through free form text, e.g., by writing "Aja did a great job—I'd work with her again." Both types of public feedback are displayed prominently within the worker's profile: public numerical scores are aggregated to a "lifetime" score as well as a "last 6 months" score, and the entire public numerical and written feedback history—including the employer's and the worker's identities—is always available to interested parties for inspection.

3.2 Private feedback before it was consequential

The platform believed that the existing public reputation system was inflated. To combat reputation inflation, the platform introduced a new, parallel reputation system that collected "private" feedback. Employers assigned private numerical feedback by rating workers on a numerical scale of 0 to 10, in answer to the question "How likely are you to recommend this [worker] to a friend or colleague?" Critically, neither other employers nor any workers can access a employer's private feedback. At first, nothing from the private feedback was used on the platform, and raters knew this.

The reason that platform was eliciting private feedback to obtain information that would help it to evaluate whether public feedback was subject to reputation inflation. The feedback forms are provided in Appendix A.

3.3 Private feedback made consequential

This feedback was kept anonymous, and, at first, was neither revealed to nor used by other market participants. After collecting private feedback for 9 months, the decided to begin releasing publicly batched aggregates of this private feedback score. This new private feedback score was displayed prominently on the profile of each worker, similar to the status-quo display of public feedback score aggregates.

The new private feedback score differed from the pubic feedback score in two important ways. First, it would only be updated after the worker received five new private feedback scores from different raters. For example, consider a worker who received private feedback scores of $1,2,\ldots 10$, in sequence. This worker would have no private feedback score at first, then once the worker received 5 private feedback ratings, the worker would have a private feedback aggregate score of 3, as $3 = (1+2+3+4+5)/5$. The feedback would stay at 3 until it changed to $5.5 = (1+2+3+\ldots 10)/10$. Second, the identity of the employers/raters would not be revealed to any other platform user, including the rated worker. Employers (raters) were also told that this score, while anonymous, would be used in the manner described above.

To the extent that employers used this new private feedback score in their hiring decisions, this change increased the workers' cost of receiving bad private feedback. In the language of our model, this change decreased $w_{s=0}$. The platform's hope was that by not allowing workers to find out which employer gave feedback, the distribution of the employers' reflection coefficients c would remain close to zero. However, if employers simply did not want to hurt the rated workers—they have altruistic concerns—then even the anonymized, batched release of private feedback scores will still increase the rater costs. The test is whether the private feedback score inflates after it becomes revealed and hence consequential.

3.4 Event study on the effects of the revelation

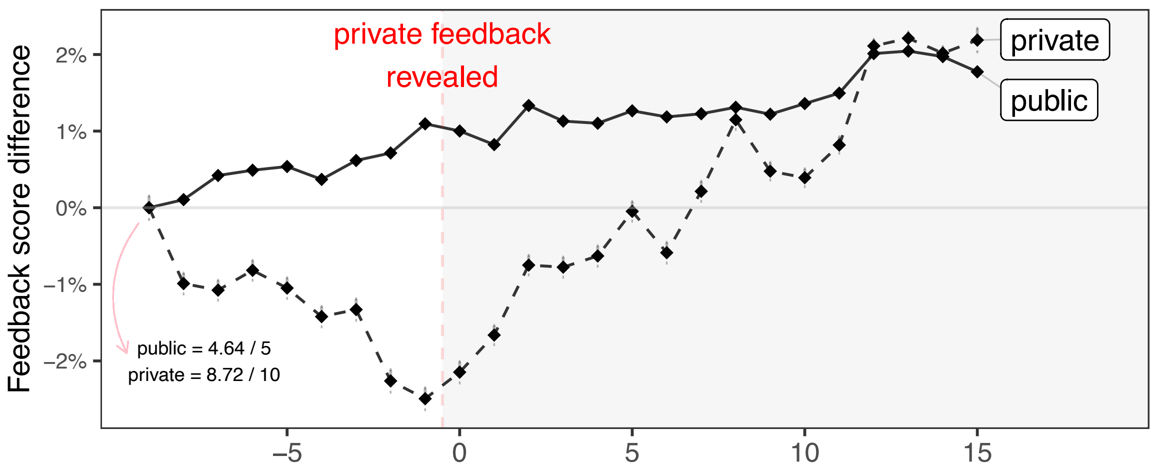
To begin, we simply plot monthly averages of feedback scores over time. Figure 2a plots both the average public and private feedback scores for transactions that received both types of feedback. We normalize scores by the first value of each respective time series, and report them as percentage changes over these values. At the start, the mean score was 4.64 for public feedback and 8.72 for private feedback. The post-revelation period is indicated by the vertical red dashed line and the gray-shaded region.

We can see in Figure 2a that before the revelation, public and private feedback scores were diverging for the same transactions. Private feedback was doing down, while public feedback was going up, continuing its long-running process of inflation.

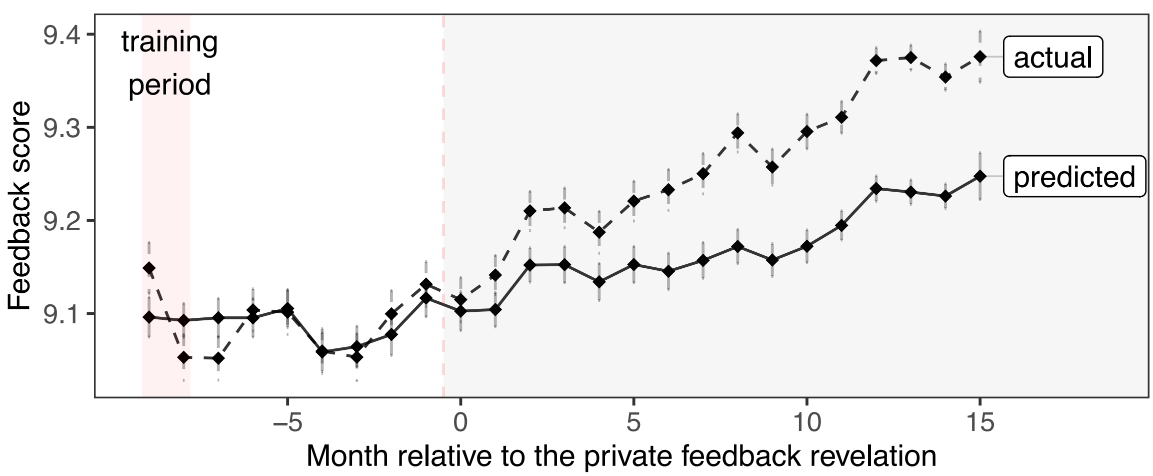
After the revelation, private feedback scores suddenly started increasing rather than declining. This break in trend is a smoking gun; the nature of the change that preceded this trend break suggests that rater altruistic cost concerns pulled the trigger. But there are plausible alternative explanations.

Figure 2: Evidence of inflation in private feedback after it was revealed publicly

(a) Public and private feedback over time for all jobs



(b) Actual and predicted private feedback over time for jobs that received written feedback



Notes: This figure shows the effects of the private feedback revelation on private feedback scores. The top panel plots monthly averages for the public scores (solid line) and the private feedback scores (dashed line) assigned by employers to workers for the same transactions. The sample includes all jobs, and the scores are normalized by the value of the first observation for each type of feedback. The bottom panel plots monthly averages for the private feedback scores (dashed line) and the predicted feedback scores (solid line) assigned by employers to workers. The sample includes all jobs that received written feedback, and the predicted scores are derived from the employers' public written feedback, by fitting a predictive model using data from the periods indicated by the red-shaded area (see Section 3.5 for more details). For both panels, the vertical dashed red line and the gray-shaded area indicate the post-revelation period.

3.5 Disentangling improvements in fundamentals from inflation

One potential explanation for the rise in the private feedback score post-revelation is that the introduction of the new feedback measure—the revelation of private feedback—improved marketplace fundamentals. For example, if the new private feedback measure helped employers make better choices or gave stronger incentives to workers, we might expect rater satisfaction and hence ratings to improve for fundamental reasons rather than inflation. Ironically, if true, this would also suggest the existing public feedback score had lost its informativness.

But one piece of evidence against this alternative explanation is that, in the post-revelation period, private feedback scores started increasing immediately, but the rate of increase in the public feedback scores did not change noticeably. Fundamental improvements that increased rater satisfaction were not showing up in the public score. However, public feedback is already inflated, and hence it might no longer be a sensitive instrument. A better approach is to use the "alternative measures" approach to decompose feedback changes.

The alternative measures approach consists of learning the expected value for a primary feedback score, which we suspect is subject to inflation, conditional upon an alternative measure of rater satisfaction. This learned conditional expectation function can then be applied to new transaction data, predicting what the primary average score "should" have been given the alternative measure. This approach allows us to net out the increase in the primary feedback measure that is not attributable to changes in marketplace fundamentals that affect rater satisfaction. In our context, the primary feedback score will be the private feedback scores, and the alternative measure will be the written feedback left by employers after each transaction. For a detailed description of the alternative measures approach and its econometric properties, see Section 4 in Filippas et al. (2022)

Figure 2b plots the monthly average private feedback and the monthly average *predicted* private feedback for transactions that received both types of feedback. The predicted private feedback is the prediction of a model trained using data from a pre-revelation period, which indicated by the red-shaded region.[3]

Prior to the revelation, the actual and predicted private feedback scores track each other closely. After the revelation, the two scores diverge substantially: the predicted feedback scores increase, but at a much lower rate than the actual private feedback scores. This suggests that the revelation had positive effects on rater satisfaction, as predicted feedback scores increased thereafter, but also that much of the increase was due to inflation. However, to the extent

[3]To learn the predictive model, we use a natural language processing pipeline. First, we preprocess the text of each review by stripping it of accents and special characters, lowercasing it, and removing stopwords. We then create a matrix of token counts (up to 3-grams), and use the TFIDF weighting method. We then conduct an grid search, evaluating each configuration of hyper-parameters for a range of predictive models, using a 5-fold cross validation in terms of average squared error. We then the fitted model with the highest prediction to estimate out-of-sample feedback scores of the written feedback for the entire sample. The training period is indicated by the red-shaded region in Figure 2b.

written feedback was also inflating in sentiment, we are likely to overestimate improvements in fundamentals.

3.6 Quantifying the effects of the revelation on private feedback scores

To quantify the effect of private feedback revelation, we switch to a difference-in-differences regression framework. We use the same actual and predicted private feedback data as in Section 3.5. It is worth noting that, unlike a difference-in-differences analysis where the treated and control units are separate entities (e.g., a treated worker subject to a higher minimum wage in some state, and a different worker in a different state with the status quo), the two kinds of feedback scores in our data are literally for the same units—namely a job contract. As such, we can do the first difference ourselves and use the difference between the numerical private feedback and the predicted private rating based on the written text, Δs, as our outcome. By taking this difference, we eliminate the need (and possibility) of including period-specific fixed effects, and so our baseline specification is simply

$$\Delta s_{it} = \beta_0 + \beta_1 \cdot \textsc{PostRevelation}_{it} + \epsilon \tag{7}$$

where PostRevelation is an indicator the contract occurred after private feedback was publicly revealed.

Table 1 reports the estimated contract-level effect of the private feedback revelation through four regression specifications. All standard errors in this table are clustered at the level of the individual employer.

Column (1) reports an estimate using Equation 7. We can see that after the switch to revelation, the gap increased, with private feedback scores becoming considerably more positive compared to the predicted score based on text sentiment. The effect size of 0.1 is about 5.4% of the population standard deviation in the feedback differences Δs.

A limitation of conducting a contract-level analysis is that we overweight employers and workers with many contracts. To address this limitation, Column (2) restricts the sample to employers and workers with at most one completed contract per-month in the pre-revelation period, and at most one completed contract per-month in post-revelation period. This restriction results in dropping about 50% of the observations, but leaves the estimated effect size similar in size.

One plausible reason for the emergence of this gap is due to selection of raters with systematically large or small gaps between public and private feedback. To assess this possibility, Column (3) adds an employer-specific fixed effect to the regression. The effect size is somewhat smaller when we include the employer fixed effect, but its magnitude remains close to the full-sample effect.

Although we have modeled the treatment as a single treatment effect, some of the treat-

Table 1: Effects of the private feedback revelation on private feedback scores

	Dependent variable:			
	Δs, (Actual - Predicted) Private Feedback			
	(1)	(2)	(3)	(4)
Post-revelation	0.100^{***}	0.110^{***}	0.094^{***}	0.036^{**}
	(0.008)	(0.008)	(0.013)	(0.017)
Post × Month				0.008^{***}
				(0.001)
Constant	-0.002	-0.061^{***}		
	(0.008)	(0.007)		
≤1 contract per period	N	Y	Y	Y
Employer FE	N	N	Y	Y
Observations	897,603	432,944	432,944	432,944
R^2	0.001	0.001	0.512	0.512

Notes: This table reports estimates of the private feedback revelation effect on private feedback scores. Column (1) reports a regressions where the outcome is the difference between the actual and the predicted private feedback scores, and the independent variable is an indicator variable for the private feedback revelation. Column (2) restricts the sample to employers and workers with at most one completed contract per-period in the pre-period, and in the post-period. Column (3) adds an employer-specific fixed effect to the regression. Column (4) adds adds a linear time trend for the post-period. All standard errors are clustered at the employer level. Significance indicators: $p \leq 0.1 : \ddagger$, $p \leq 0.05 : *$, $p \leq 0.01 : **$, and $p \leq .001 : ***$.

ment effect detected in Columns (1)-(3) was likely the accumulation of an increasing gap in the post-period. See Appendix B for a visual analysis indicating a trend break. As such, for our last specification, we include in Column (4) both a post-revelation indicator and a linear time trend for the post-period, maintaining the sample restriction and the employer-specific fixed effect. We see evidence that the gap is growing over time, with a highly significant coefficient on the Post × Month term. We might expect the growth to slow, particularly as it nears the top value. Nevertheless, if we project the estimated trend into the future, the average private feedback score would reach its top value in $(10-9.37)/(0.008)/12 \approx 6.5$ years. In short, the benefits of the private feedback score are likely to diminish in the long-run: in about six-and-a-half years, we would expect little information to be left in the private feedback, though it might still be useful to detect "terrible" outcomes.

4 Discussion

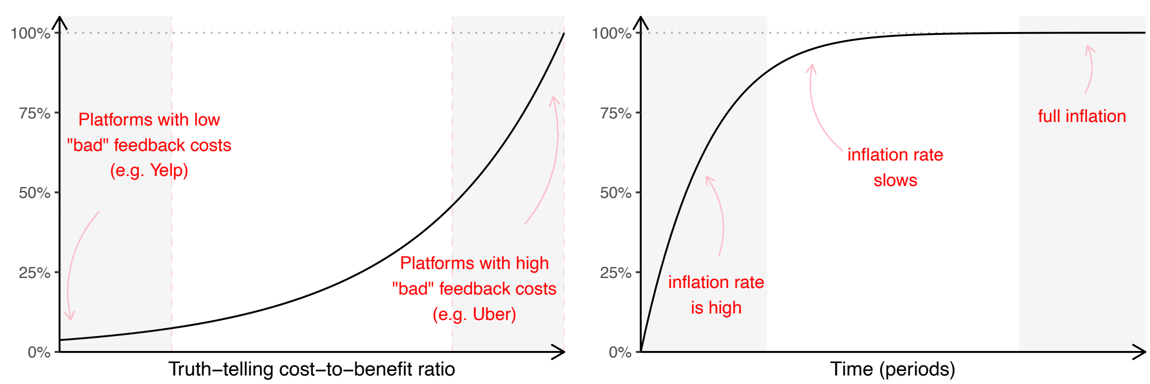
The empirical results suggest that altruistic concerns by raters are sufficient to cause reputation inflation, at least in our setting. A natural question is how well this result generalizes.

Our model predicts that reputation inflation will be acute when the ratees' costs from receiving bad feedback are high, or when the raters' reflection coefficients are high. The wage penalties from receiving bad feedback are substantial in both peer-to-peer markets and schools (Babcock, 2010; Cabral and Hortaçsu, 2010) and these ratings are highly personal, with an individual being the target of the rating.[4] It seems likely that in these settings raters will be relatively less truthful, and reputation inflation will be severe unless checked. In contrast, raters' reflection coefficients are likely smaller when assigning feedback scores to products, such as movies and restaurants, and hence reputation inflation will be less acute. Institutional ratings, such as BBB and health inspection scores, are also less prone inflation. In these cases, raters also likely view themselves as performing a service for fellow consumers, and value being known for good, honest reviews, increasing b.

To provide a graphical depiction of this intuition, we plot in Figure 3a the equilibrium inflation for different truth-telling cost-to-benefit ratios. We depict the case of $q_H = 0.8, q_L = 0.6, \alpha = 0.2, u_H = 1, u_L = 0, b = 1$, and $F \sim Exp(\lambda)$, varying the parameter λ to vary truth-telling costs. For low cost-to-benefit ratios, most raters report their feedback truthfully in equilibrium. As the cost-to-benefit ratio increases, the equilibrium inflation approaches one, and the average feedback scores become more inflated: this is the case for platforms such as Uber or Airbnb, where the reflected costs are high.

Figure 3: Properties of the reputation inflation equilibrium

(a) Equilibrium inflation and truth-telling costs (b) Inflation over time



Notes: This figure shows properties of the reputation inflation equilibrium. The left panel plots the equilibrium inflation p^* (y-axis) as a function of the truth-telling cost-to-benefit ratio (x-axis). It depicts the case of $q_H = 0.8, q_L = 0.6, \alpha = 0.2, u_H = 1, u_L = 0, b = 1$, and $F \sim Exp(\lambda)$. To vary truth-telling costs, we vary the parameter λ of the distribution of the reflection coefficients, keeping other parameters fixed. The right panel shows an example of the equilibrium convergence process, by plotting the inflation (y-axis) over time (x-axis). It maintains the same parameterization as the left panel, and fixes $\lambda = 1$. For more details on the convergence process, see the discussion in Section 4.1.

[4] In online marketplaces, ratees' costs from receiving negative feedback include that feedback scores are often the sole signal of quality, and that ratees are typically highly substitutable. On the rater side, reflection coefficients are higher because raters may be averse to harming a ratee's future prospects, are more likely to receive complaints, and because would-be trading partners may be unwilling to transact with someone who has a "strict rater" reputation.

4.1 The reputation inflation process

As in any static model, our model simply predicts the resulting equilibrium from a given set of parameters, with the model silent about the movement towards that equilibrium. However, we can introduce some simple dynamics and explore how reputations would inflate over time. Consider again a market where employers match randomly with workers in each period $t \in \mathbb{N}$, and assume that all employers start off being truthful, that is, $p_0 = 0$. In addition, assume that employers observe p_{t-1} before matches form in period t.[5]

In period 1, employers observe the current rate of inflation, $p_0 = 0$, employer-worker matches form, and employers offer wages $w_{s=0}(p_0)$ and $w_{s=1}(p_0)$, as appropriate. A fraction $n_1 = \frac{1}{2}(1 - q_L - \alpha)$ of the employers will receive borderline outputs from the workers they matched with. Among those employers who received a borderline output, a fraction $m_1 = 1 - F(bw_{s=0}(p_1))$ will lie and inflate their ratings, assigning $s = 1$ instead $s = 0$. This will push reputation inflation from $p_0 = 0$ to $p_1 = n_1 m_1$.

Because $p_1 > p_0$, we get $w_{s=0}(p_1) < w_{s=0}(p_0)$, that is, receiving bad feedback becomes costlier for workers. The has two implications for the feedback assigned after the second round's matches. First, employers who already inflated their feedback will again choose to inflate their feedback if they receive a bad product after the second round. Second, some supra-marginal employers who would have reported feedback truthfully in the first round, will instead lie and inflate their feedback after the second round. This "ratcheting down" of rater truthfulness will carry on until the process converges to an equilibrium.

Figure 3b shows an example of the inflation process that we described. We maintain the parameterization used in Figure 3a, and fix $\lambda = 1$. Starting from a zero-inflation state, employers begin inflating their feedback fast. The rate of inflation then decreases, and eventually equilibrium is reached. It is worth noting that market design changes which increase the cost of receiving bad feedback, such as the private feedback revelation that we examined in Section 3, will also have the same effect of kickstarting the reputation inflation process, as the system moves to a new equilibrium.

The convergence process we described is simple, but it is qualitatively similar to what is commonly observed in practice (Filippas et al., 2022). It is also worth noting that, although the inflation is equal to one in the equilibrium of our example, employers will continue generating bad feedback when they receive terrible output. This will result in the discrete analogue of a J-shaped rating distribution—also a recurrent empirical finding.

[5] Although employers do not observe the inflation rate directly in real-life markets, they observe proxies, such as average ratings in search rankings or worker profiles.

5 Conclusion

We showed through a simple model that raters assign inflated feedback if they incur "reflected" costs commensurate with the ratees' costs of receiving feedback. We assessed our model empirically through a quasi-experiment where a sudden increase in the ratees' feedback costs caused raters to start inflating their feedback. Because raters remained anonymous, our data suggests that rater altruism is a sufficient cause for reputation inflation.

Whether there are effective platform design responses to reputation inflation is an open question. Platforms could emphasize reviewers as performing a service for fellow consumers, thus shifting altruism, or provide other incentives for honest reviews. For example, Yelp employs mechanisms such as badges for top reviewers, and makes the feedback score distribution of each reviewer publicly accessible. Some non-digital reputation systems attempt to tackle reputation inflation directly, by imposing mandatory grading curves or stack rankings. However, it is challenging to force a distribution in settings where buyers evaluate sellers as a "flow," that is, continuously.

Platforms already take steps to lower raters' costs. These steps, such as simultaneously-revealed ratings (in place since the start of the platform we studied) and anonymizing ratings through aggregation (as was the case with the private feedback revelation that we examined), did not prevent inflation from occurring in our data. However, the model suggests they might work to slow inflation and that might be sufficient over a short enough horizon.

Our findings suggest that raters' costs come from the harm they impose on the ratees. As the potential to harm is what makes ratings effective, tackling reputation inflation is fundamentally challenging. Addressing it fully likely calls for a fundamentally different approach to reputation system design. For example, future work could examine whether it is possible to reduce the "punishment" role of bad feedback, and increase its "voice" role in helping ratees improve, or giving them other incentives for quality.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond, "Preferences for truth-telling," *Econometrica*, 2019, *87* (4), 1115–1153.

Agrawal, Ajay, John J Horton, Nicola Lacetera, and Elizabeth Lyons, "Digitization and the contract labor market: A research agenda," in "Economic analysis of the digital economy," University of Chicago Press, 2015, pp. 219–250.

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.

Bolton, Gary, Ben Greiner, and Axel Ockenfels, "Engineering trust: Reciprocity in the production of reputation information," *Management Science*, 2013, *59* (2), 265–285.

Bolton, Gary E. and Axel Ockenfels, "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, March 2000, *90* (1), 166–193.

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.

Einav, Liran, Chiara Farronato, and Jonathan Levin, "Peer-to-peer markets," *Annual Review of Economics*, 2016, *8*, 615–635.

Filippas, Apostolos, John J Horton, and Richard J Zeckhauser, "Owning, using, and renting: Some simple economics of the âĂIJsharing economyâĂİ," *Management Science*, 2020, *66* (9), 4152–4172.

_, John Joseph Horton, and Joseph Golden, "Reputation inflation," *Marketing Science*, 2022.

Fradkin, Andrey, Elena Grewal, and David Holtz, "Reciprocity and unveiling in two-sided reputation systems: Evidence from an experiment on Airbnb," *Marketing Science*, 2021, *40* (6), 1013–1029.

Gutt, Dominik, Jürgen Neumann, Steffen Zimmermann, Dennis Kundisch, and Jianqing Chen, "Design of review systems–A strategic instrument to shape online reviewing behavior and economic outcomes," *The Journal of Strategic Information Systems*, 2019, *28* (2), 104–117.

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

_ , William R Kerr, and Christopher Stanton, "Digital labor markets and global talent flows," Technical Report, National Bureau of Economic Research 2017.

Kokkodis, Marios, "Dynamic, multidimensional, and skillset-specific reputation systems for online work," *Information Systems Research*, 2021, *32* (3), 688–712.

Li, Xinxin and Lorin M Hitt, "Self-selection and information role of online product reviews," *Information Systems Research*, 2008, *19* (4), 456–474.

Rojstaczer, Stuart and Christopher Healy, "Where A is ordinary: The evolution of American college and university grading, 1940-2009," *Teachers College Record*, 2012, *114* (7), 1–23.

Tadelis, Steven, "Reputation and feedback systems in online platform markets," *Annual Review of Economics*, 2016, *8*, 321–340.

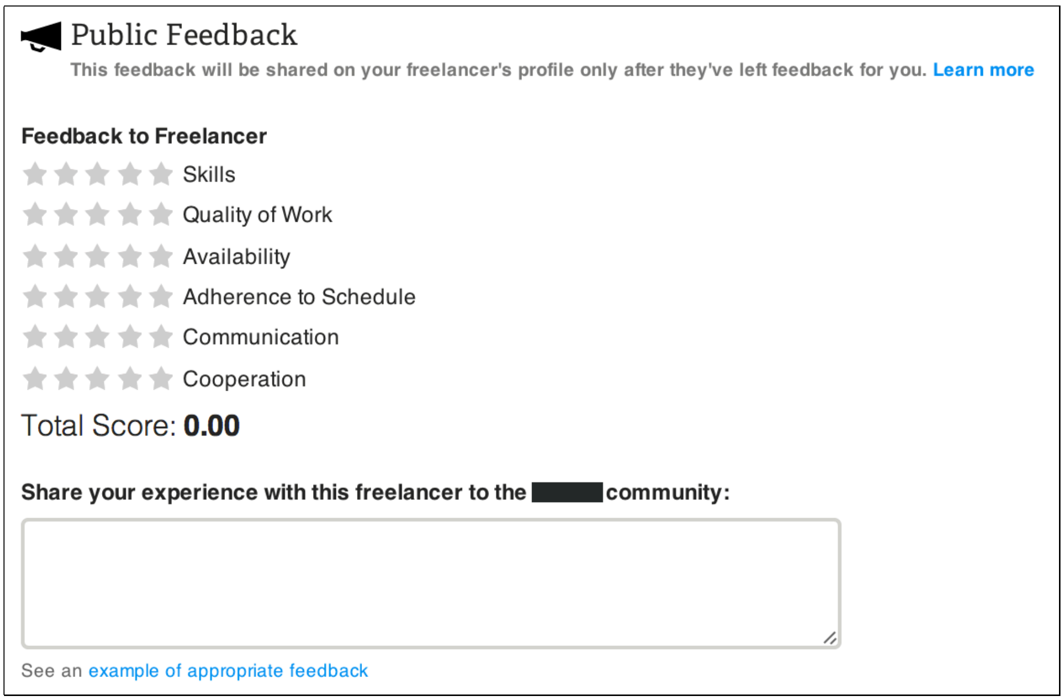
Yang, Mingwen, Zhiqiang Zheng, and Vijay Mookerjee, "The race for online reputation: Implications for platforms, firms, and consumers," *Information systems research*, 2021, *32* (4), 1262–1280.

A Feedback forms

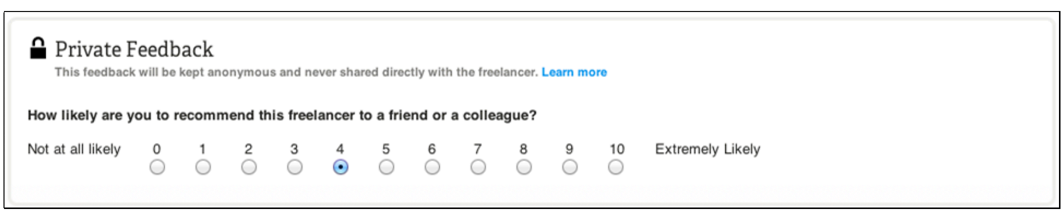
Figure 4 shows the public and private feedback interfaces for employer-to-worker feedback. Numerical feedback is elicited on a 1-5 scale across six weighted dimensions, which are aggregated to a total score. Written feedback is elicited as free-form text. Private feedback is elicited as numerically. The two interfaces are displayed on the same page, and hence employers are asked to assign both public and private feedback for the same transactions.

Figure 4: Post-transaction feedback form

(a) Public feedback interface.



(b) Private feedback interface.

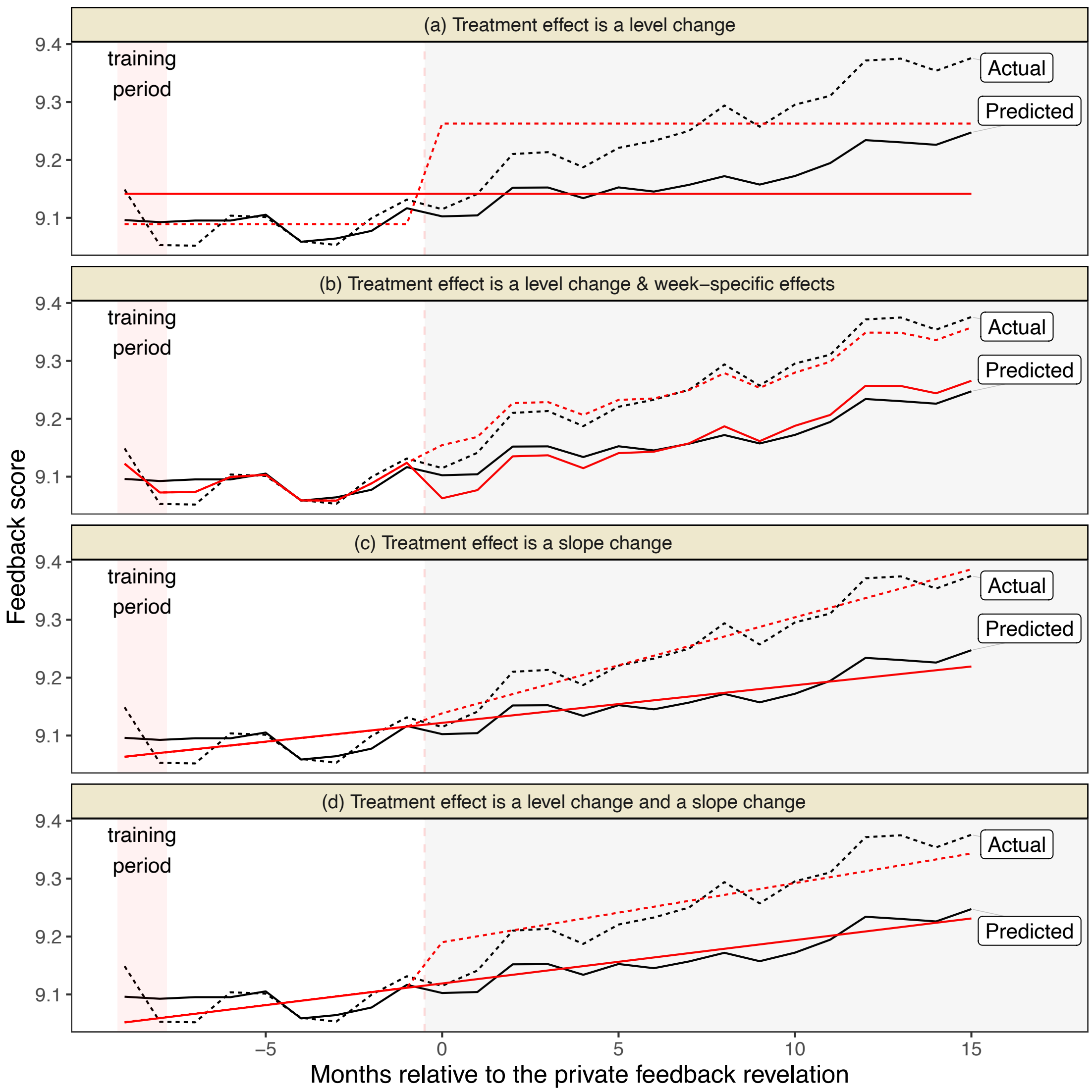


Notes: This figure shows the feedback form that employers were asked to fill upon completion of their contracts with workers. The top panel shows the public feedback elicited, and the bottom form shows the private feedback elicited. The two types of feedback were elicited on the same page.

B Aggregate effects of the private feedback revelation

In this section, we examine the aggregate effects of the private feedback revelation simply and visually. Figure 5 plots the estimated aggregate-level treatment effects, using per-period averages for each type of feedback. As we are afforded some flexibility over the choice of specification, we plot the results of four alternative specifications.

Figure 5: Monthly average actual and predicted private feedback scores



Notes: This figure shows the aggregate effects of the private feedback revelation on private feedback scores. In each panel, the black lines plot the monthly average private (solid line) and predicted private (dashed line) feedback scores assigned by employers to workers. The red lines plot predictions from four difference-in-differences specifications, described in the strip text of each panel. The vertical dashed red line and the gray-shaded area indicate the post-revelation period. The sample includes jobs that received public written feedback, and the predicted scores are derived from the employers' public written feedback, with the predictive model fit using data from the periods indicated by the red-shaded area.

Panel (a) reports the simplest specification, where the treatment is allowed to only have a level effect, and the two feedback-types are allowed to differ by a fixed amount in the pre-period. This specification fails to capture the underlying time trend in both series, and especially for the actual feedback in the post period.

Panel (b) maintains the assumption of a level effect, and includes a week-specific effect. This specification captures better 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 on and under-estimating it later on—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.

Panel (c) gives both types of feedback a common linear time trend, but then allows that trend to change in the post-period for the actual private feedback. With a common slope, the fit in the pre-period improves substantially.

Panel (d) allows for both a level treatment effect and a change in slopes, and results in little change. The last two specifications seem to work best, with the predicted series closely matching the realized values. We make use of this insight in Section 3.6, where we estimate the effects of public revelation at the level of the individual contract, rather than at the level of monthly averages.