The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment*

Chris Nosko†

Steven Tadelis‡

February 8, 2015

Abstract

Reputation mechanisms used by platform markets suffer from two problems. First, buyers may draw conclusions about the quality of the platform from single transactions, causing a reputational externality. Second, reputation measures may be coarse or biased, preventing buyers from making proper inferences. We document these problems using eBay data and claim that platforms can benefit from identifying and promoting higher quality sellers. Using an unobservable measure of seller quality we demonstrate the benefits of our approach through a large-scale controlled experiment. Highlighting the importance of reputational externalities, we chart an agenda that aims to create more realistic models of platform markets. *JEL* D47, D82, L15, L21, L86

*We are grateful to many employees and executives at eBay without whom this research could not have been possible. We thank Dominic Coey, Andrei Hagiu, Ali Hortaçsu, Dimitriy Masterov, Carl Mela, and Glen Weyl for very helpful comments.

†University of Chicago Booth School of Business and eBay Research Labs. Email: cnosko@chicagobooth.edu

‡UC Berkeley Haas School of Business, NBER and eBay Research Labs. Email: stadelis@berkeley.edu

1 Introduction

Decentralized marketplaces constitute some of the most fundamental building blocks of economic activity. eBay, one of the first internet success stories, morphed from a used-goods auction site into one of the largest platform markets with over sixty billion dollars of merchandise traded in 2013. Other prominent online platform markets include Uber, Amazon's Marketplaces, AirB&B, and Taobao to name a few. These marketplaces use some sort of decentralized "reputation" mechanism to mitigate the concerns about market failures that result from asymmetric information due to the anonymity of traders.

In this paper we argue that decentralized sellers in platform markets do not internalize the impact of their actions on the marketplace as a whole. In particular, one disappointing transaction may cause a buyer to update his beliefs about the quality of *all* sellers on the platform, resulting in a *reputational externality* across sellers. Furthermore, if some buyers leave the platform after a disappointing transaction without leaving feedback then the platform's reputation mechanism will be positively biased and therefore less effective.

We proceed to study the challenges faced by market platforms in the presence of reputational externalities and biased feedback. We explore the limits of reputation mechanisms in the face of these problems, their impact on the marketplace, and ways in which a platform designer can mitigate these adverse impacts. As such, our paper offers three contributions to the literatures on market design and on reputation mechanisms.

First, using data from eBay that records the *actual* behavior of buyers, we show that buyers respond to low quality transactions by choosing to leave the platform. To do this we exploit the bias in feedback to create a new measure of seller quality. Second, we establish that reputational externalities exist, and that feedback is biased. Last but not least, we propose a mechanism to mitigate the externality problem in which good-quality sellers are prioritized in search results. We conduct a field experiment where we change search results for a randomly chosen subset of buyers using our measure of quality and find that this approach increases the quality of transactions and, consequently, the retention of buyers.

We begin our analysis by suggesting a simple conceptual framework of buyer behavior in online marketplaces. We then construct a longitudinal dataset using eBay transactions that follow a cohort of new buyers who joined eBay anytime during 2011 and include all their transactions through May 2014. The data include every transaction made by this cohort, including item characteristics and the item's seller. The goal is to measure how the quality of a transaction affects the future behavior of buyers on the platform.

We first establish that the standard measure of a seller's quality, his reputation feedback, is highly skewed and omits valuable information. The "percent positive" (PP) measure for each seller is computed by dividing the number of transactions with positive feedback by the number of transactions with any feedback for that seller. In our dataset, PP has a mean of 99.3% and a median of 100%, consistent with other studies that use eBay data (Dellarocas and Wood, 2008). Hence, a central challenge is to construct a measure that more accurately reflects a seller's true quality. We construct a new quality measure that we call "effective percent positive" (EPP), which has a mean of 64% and a median of 67%, and exhibits significantly more variability than PP. Most importantly, because EPP is not observable, buyers cannot select on it and we can interpret the results of our analysis as causal.

Our conceptual framework guides our empirical analysis of the actual behavior of buyers with respect to how current transactions affect their future behavior. In particular, we use a "revealed preference" approach to study the effect of a seller's EPP in a *current* transaction on the buyer's propensity to *continue* buying on eBay. This distinguishes our paper from a long list of papers that collect scraped data from marketplaces and are limited to consider only prices and quantities. This approach allows us to get to the heart of the question of whether reputation mechanisms are indeed steering buyers away from low quality sellers.

As our framework suggests, a buyer who has a better (higher EPP) experience on eBay will be more likely to continue to transact on eBay again in the future. Furthermore, Bayesian buyers with more experience should be less sensitive to their current transaction quality. That is, the response to a negative experience early in a buyer's tenure on eBay should be more severe than later in his purchasing sequence. We confirm the first hypothesis using the

data and demonstrate that EPP is a useful measure of seller quality. Patterns in the data offer support for the second hypothesis as well.

To establish the effect of improving buyer experiences by prioritizing higher quality sellers, we report results from a controlled experiment on eBay that incorporated EPP into eBay's search-ranking algorithm. The experimental approach alleviates any concerns about selection and endogeneity between buyers and a seller's EPP. We selected a random sample of eBay buyers who, when searching for goods on eBay, were shown a list of products that promoted seller EPP compared to a control group in which this was not done. The results confirm the conclusions from the regression analyses of the cohort data described above and show that treated buyers who were exposed to higher EPP sellers were significantly more likely to return and purchase again on eBay compared to the control group of buyers. Combining the experimental data with information on a buyer's experience on eBay up to the time of the experiment, suggests that buyers act as if they are Bayesian learners.

A growing empirical literature explores the effect of online feedback on market outcomes (Luca, 2014) and strategic reasons for biased feedback (Mayzlin et al., 2014). Several papers have focused attention on eBay's reputation system, including Bajari and Hortaçsu (2004), Bolton et al. (2013), Cabral and Hortaçsu (2010), Dellarocas (2003) and Klein et al. (2013) to name a few. Within this literature, the paper closest to ours is Dellarocas and Wood (2008) who reveal the problem of skewed feedback and propose an econometric method to uncover a seller's actual reputation.[1] We distinguish our work by focusing attention on reputational externalities and on the extent to which the observable reputation measures are biased compared to actual seller quality. More important, unlike most studies that analyze the effects of reputation on outcomes such as prices and probabilities of sale, we use buyers' revealed preferences from their decisions over whether or not to exit the platform to focus on a different set of questions.

[1]Their method relies on the historical two-way feature of eBay's reputation mechanism. However, in 2008 eBay changed the mechanism so that buyers can no longer receive negative feedback, implying that their mechanism can no longer be used on eBay data.

Properly construed, buyer exit maps closely into transaction satisfaction and hence provides a measure of overall platform welfare. Instead of looking for evidence that the reputation system has *some* effect on outcomes regardless of how this maps into welfare, we provide empirical evidence that platforms may be far away from the optimum, a wedge created by reputational externalities on the seller side of the market.[2]

A large theoretical literature argues that reputation mechanisms mitigate inefficiencies in markets with asymmetric information (see Bar-Isaac and Tadelis (2008) for a survey). By publicly revealing ratings from past transactions, sellers are punished for delivering bad quality through the loss of future business from other market participants. Such mechanisms have been credited with sustaining markets such as long distance trade during the Middle Ages (Greif, 1989) and are cited as reasons that online anonymous markets were able to come into existence in the first place. Within this large theoretical literature, Tirole (1996) stands out as distinguishing individual reputations from group reputations and proposes a theoretical model of collective reputations in which a group's reputation aggregates the reputation of its members. His focus, however, is on reputation persistence and the way in which the incentives of group members are influenced by the group's reputation, instead of the extrernality problem that we emphasize. In our view, the problem of reputational externalities extends beyond market platforms, and are relevant for any setting in which an organization's reputation as a whole is influenced by the experiences that its clients have with individuals in the organization.

Equipped with a more accurate measure of quality, we explore the extent of reputational externalities as well as how a platform should intervene and use levers other than the reputation system to increase platform quality. Two extreme mechanisms can be used by platforms to manage seller-quality on their sites. At a draconian extreme, the platform can

[2]Consider, for instance, the well-documented stylized fact that sellers on eBay with higher reputational measures receive higher prices for their auctioned items (e.g., Cabral and Hortaçsu (2010)). This may be a sign that the reputation system works well in separating out seller types, but it may also be a sign that large pools of naive buyers think they are getting a "good deal" and enter into transactions that they are deeply unhappy with. Thus, the claim that "eBay's impressive commercial success seems to indicate that its feedback mechanism has succeeded in achieving its primary objective (Dellarocas (2003))" side steps the questions we are principally interested in, namely, given imperfect buyer information, how much *more* successful could a platform like eBay be.

expel any seller once it learns that a transaction was less than perfect, resulting in a small but selected pool of high quality sellers. On the other more "laissez-faire" extreme, the platform can provide a reputation system and let buyers choose whom to buy from while managing their own risk. As a platform shifts from a more draconian to a more laissez-faire approach, it trades off quality in favor of variety. Historically, eBay had been closer to the laissez-faire extreme, but has recently been taking a more active role in managing the marketplace. Hui et al. (2014) investigate the role of some of these changes.[3]

Based on our analysis, we advocate that online marketplace platforms use search technology to control buyer experience, establishing a middle-ground between transaction quality and seller variety. Our approach rests on the platform's ability to use its search algorithm for controlling the exposure of seller quality to buyers. To the best of our knowledge, the experimental evidence we present is some of the first to show how buyers respond to truly exogenous shifts in search rankings. We conclude by discussing a more general agenda for studying reputation and quality in the design of platform markets.

2 Conceptual Framework

We wish to distinguish between two possible scenarios for a marketplace platform. The first is that buyers see the platform as a means of gaining access to sellers, but they neither consider characteristics of the platform itself, nor do they believe that sellers on the other side of the platform represent the platform as a whole. In this case, there are no externalities across sellers. A buyer updates *only* on the quality of the seller that he interacted with. If the

[3]These measures include actively seeking to weed out bad quality sellers and creating a "buyer protection" program that allows buyers to voice complaints to the platform directly about a transaction for potential reimbursement, rather than having to go through the individual sellers. See, for example, "eBay to Get More Involved in Transaction Disputes", http://www.pcworld.com/article/163099/article.html. In contrast to eBay, Amazon has always extensively pruned its seller pool on Amazon Marketplace, making it more difficult to join, holding transaction receipts in escrow, and removing sellers swiftly. Similarly, Stubhub (an eBay company) has been much more careful to control the buyer experience. Buyers purchasing on Stubhub are not aware of which seller they are purchasing from and all disputes are handled with Stubhub directly. These policies completely negate the need for a reputation system and essentially mean there are no externalities across sellers as they are internalized by the platform.

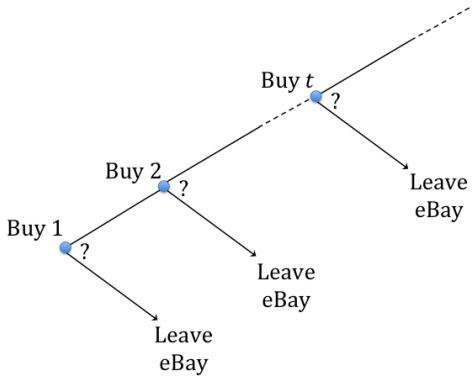

Figure 1: A Buyer's Dynamic Bayesian Decision Problem

transaction goes badly then he may not deal with that seller again, but this does not affect the buyer's willingness to transact with other sellers on the platform.

In the second scenario the buyer uses outcomes of individual transactions to form beliefs about the whole platform. To consider this, imagine a dynamic Bayesian decision problem of a buyer who arrives at the marketplace platform for the first time and is contemplating whether or not to purchase an item. His decision to purchase will rely on three basic elements: first, how much he enjoys the site experience; second, what are his expectations about the quality of the transaction; last, conditional on his belief, how price competitive is the site compared to other comparable marketplaces. If he decides to purchase, then after he receives the item he will update his beliefs about the quality of the site, and decide whether or not to purchase again, and so on, as depicted in Figure 1.

Buyers can use a seller's past performance to form expectations about the quality of the seller, and by association, the marketplace overall. Every time the buyer makes a purchase, he collects an observation through which he updates his prior belief about the site's and the seller's expected quality. If the experiences were bad enough, he will update his belief about quality downward enough so as to decide to leave the platform altogether. If, however, his experience was good, he will update his posterior in a positive way and continue to purchase from other sellers on the marketplace platform.

This framework of Bayesian updating also implies that the more transactions a buyer has made, the tighter will be his posterior, and this in turn implies that the effect of early experiences will be much more influential on the next purchase decision than later experiences.[4] It follows, therefore, that if a buyer experiences a relatively bad transaction earlier in the dynamic decision problem, then he is more likely to leave the marketplace than if he experiences the same quality transaction after several good experiences. This simple observation will form the basis for the analysis on buyer behavior in Section 5.

3 Reputation and Transaction Quality at eBay

eBay's reputation mechanism is often described as a resounding success for two reasons. First, eBay exists as a successful business despite the complete anonymity of the marketplace. Second, many observable reputation characteristics correlate with our prior notions of the directional movement that these measures should induce. For instance, sellers with higher reputation scores and more transactions receive higher prices for their products. Similarly, reputation seems to matter more for higher priced goods than for lower priced goods.[5]

When buyers complete a transaction on eBay, they can leave either a positive, negative, or neutral feedback score, or leave no feedback at all. About 65% percent of buyers leave feedback on a transaction and eBay uses this information to provide several observable seller reputation measures. The first, *percent positive* (PP), is defined as the seller's number of positive feedbacks divided by the sum of his number of positives, neutrals and negatives.[6] The second, *feedback score*, is a summed value of the number of positive feedbacks minus the number of negative feedbacks. The third is a badge that certifies a seller as an "eBay Top Rated Seller" (ETRS). This designation is bestowed on sellers that meet a series of criteria

[4]This heuristic framework can easily be formalized using a standard dynamic model of a Bayesian decision maker that faces a distribution of quality with a well defined prior on the distribution of quality. Due to the well-understood nature of this dynamic problem, it would be redundant to offer the formal model.

[5]See Bajari and Hortaçsu (2004) and Cabral and Hortaçsu (2010) for more on these facts.

[6]To be precise, these numbers only look back at the last 12 months of a transaction for a seller and exclude repeat feedback from the same buyer for purchases done within the same calendar week.

believed by eBay to be an indication of a high quality seller.[7] All of these measures are displayed when a user views detailed item information. Figure 2 shows these measures for two different sellers. Seller A has a percent positive score of 96.9 and a seller feedback score of 317, while seller B has a percent positive of 99.5, a seller feedback score of 44949, and is a part of the ETRS program (indicated by the badge that reads "Top Rated Plus").

Two obvious problems exist for using the observable reputation measures to examine the effect of seller quality on future outcomes. First, buyers select the sellers they purchase from, leading to a bias in estimates of long term benefits from interacting with higher quality sellers. For instance, a frequent eBay user – one that is likely to return to the site – may be more willing to take a chance on an observably low quality seller if there are compensating differentials. Indeed this is exactly what we found in our analyses described below. If we simply used the observable measure of seller quality, we might incorrectly conclude that interacting with an observably low quality seller causes a buyer to come back in the future. Furthermore, in equilibrium, observably lower quality sellers might adjust other features of their offering, such as the price, to compensate for their lower observable feedback measures. This adjustment, if not completely controlled for, would also lead to a bias in our estimates.

Second, the feedback measures are highly skewed. Figure 3 displays the histogram of seller PP from the dataset described in detail in the next section. The X-axis starts at 98%, which is the tenth percentile, and the median seller has a score of 100%. This could be indicative of a reputation system that works extremely well – bad sellers exit when their score falls even slightly, leading to a high positive selection. Unfortunately, this is not the case. Out of over 44 million transactions completed in October of 2011 on eBay's U.S. marketplace, only 0.39% had negative feedback, while at the same time, over 1% had an actual dispute ticket opened, a step that takes substantially more effort on a buyer's part than leaving negative feedback. Furthermore, over 3.3% of email messages from buyers to sellers sent *after* the transaction

[7]See Hui et al. (2014) for a lengthy discussion of this program. We exclude from the analysis a separate set of seller ratings, called the "detailed seller ratings," which give buyers the opportunity to rate the seller at a finer-grained level. We do this because based on analysis of the internal eBay clickstream logs, fewer than 1 percent of buyers ever click on the page that contains this information. Because of this, we restrict analysis to data contained on the main view item page.

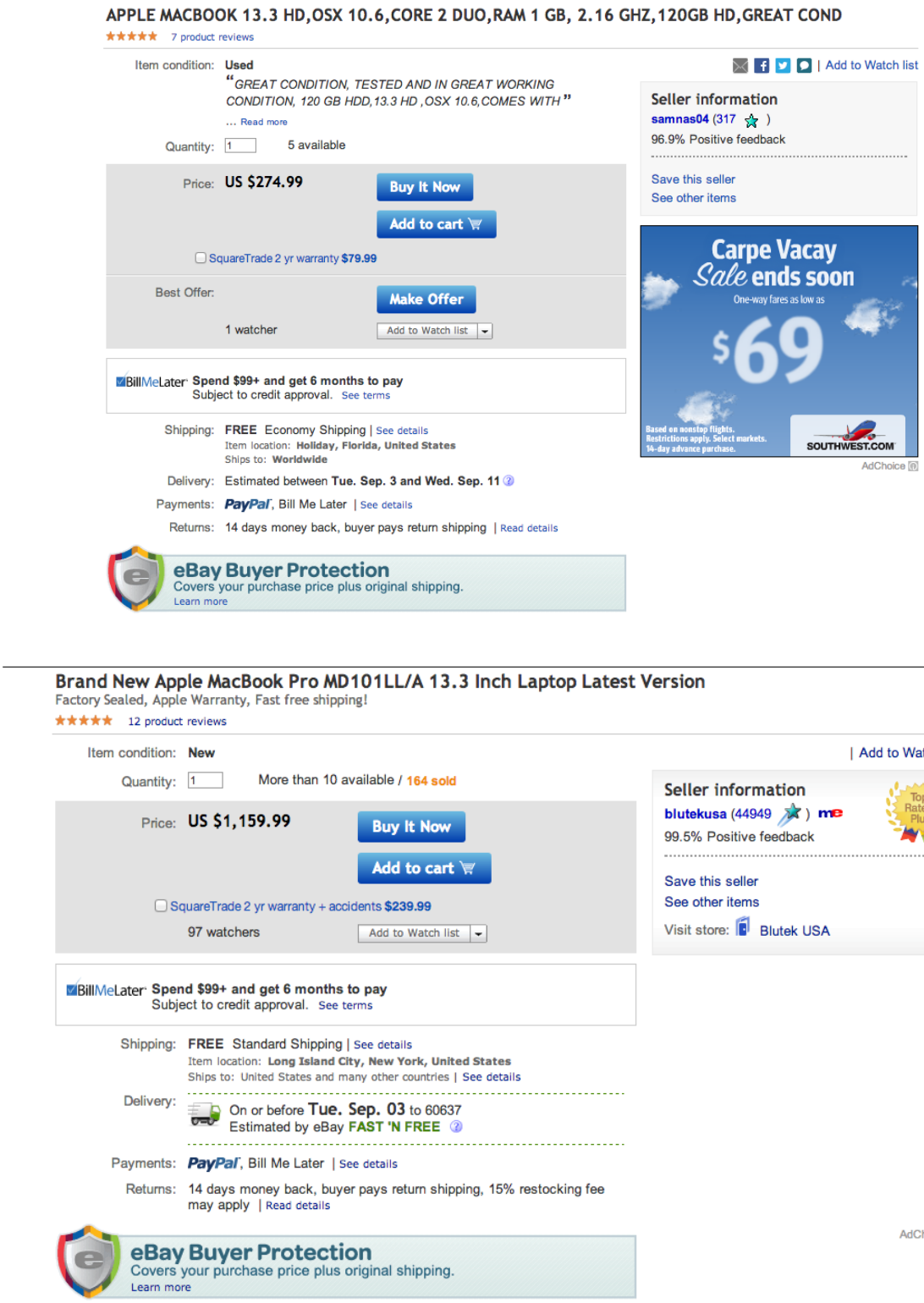

Figure 2: Seller reputation information as displayed to buyers

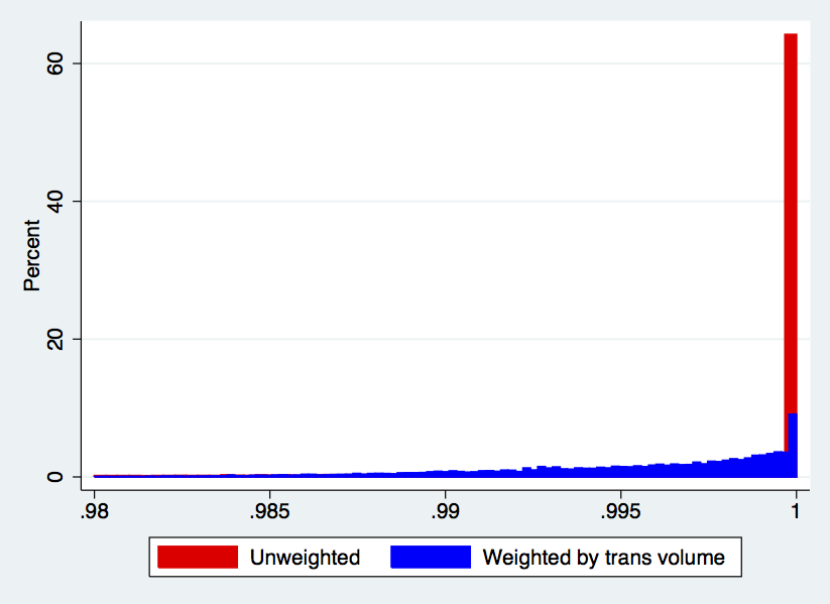

Figure 3: Percent Positive of Sellers

concludes include language that implies a bad buyer experience. (See Masterov et al. (2014) for more on this measure.) This indicates that there are a substantial number of transactions that went badly for which negative feedback was not left.[8]

Another potential problem is that many buyers may have trouble interpreting the numbers they are presented with. Naively, one may think that a score of 98% is excellent (in some sort of absolute scale). In reality, a score of 98% places a seller below the tenth percentile of the distribution. Figure 4 plots the distribution of seller feedback scores. Seller feedback scores are widely dispersed and possibly hard to interpret. How should buyers interpret a number like 161 (the median) and how should buyers form expectations about the service level they will receive from sellers with different feedback scores? These interpretation problems may be especially pronounced for new users who may not have seen enough sellers to be able to judge the scale accurately, a point to which we return below.

[8]This in fact proves the central conjecture in Dellarocas and Wood (2008) who claim that silence in feedback includes many transactions for which buyers had bad experiences but chose not to report them.

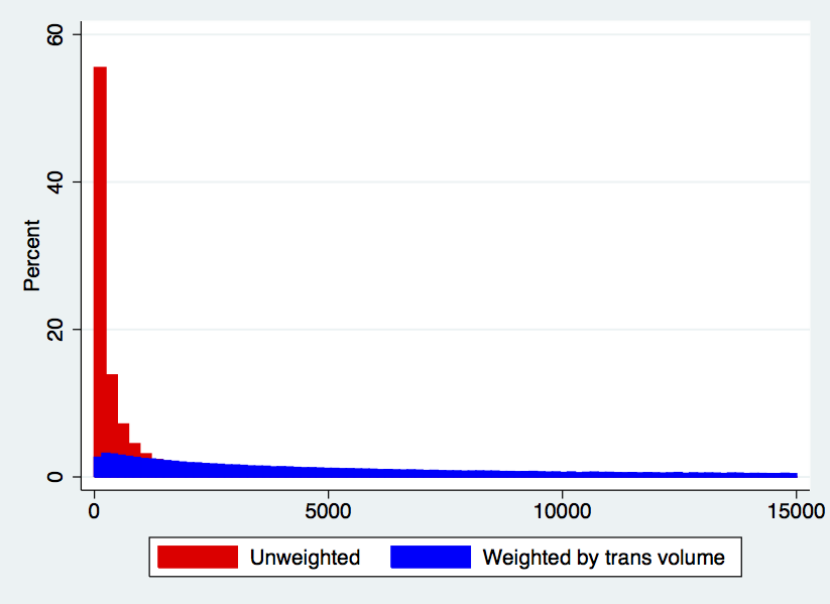

Figure 4: Seller Feedback Scores

Buyers may choose not to leave negative feedback because it is not anonymous and sellers historically reacted by reciprocating.[9] Anecdotal evidence shows that sellers sometimes react badly to negative feedback, harassing buyers in an attempt to get them to change it.[10] If it is more costly to leave negative rather than positive feedback then feedback will be biased.[11]

We proceed to construct a measure of *unobservable* seller quality based on the idea that buyers who experience a bad or mediocre transaction are likely to be silent and not leave any feedback at all. If this is the case, then silence will disproportionately indicate worse transactions. To operationalize this, we measure the propensity of positive feedback for any

[9]Up until 2008, both parties could leave negative feedback, and after that sellers can only leave positive feedback or no feedback. There is a long history of reciprocal feedback behavior before the 2008 change as documented by Bolton et al. (2013).

[10]In one case, a seller called the buyer and threatened him after his negative feedback. ("eBay Shopper Says He Was Harassed By Seller," http://www.thedenverchannel.com/lifestyle/technology/eBay-shopper-says-he-was-harassed-by-seller). In another case, a buyer was sued for leaving negative feedback ("eBay buyer sued for defamation after leaving negative feedback on auction site," http://www.dailymail.co.uk/news/article-1265490/eBay-buyer-sued-defamation-leaving-negative-feedback-auction-site.html.)

[11]For example, consider a set-up in which there is a distribution of "public mindedness" among individuals that compels them to enjoy leaving feedback for the benefit of future buyers. If the costs and benefits of leaving feedback would not depend on the quality of the transaction, then the feedback left should be unbiased. However, if the cost of leaving truthful feedback is higher for bad transactions due to the harassment costs, then such a skew in feedback will result.

given seller. Controlling for observable feedback measures (PP, feedback score, and seller standards), we conjecture that a seller with a lower propensity of positive feedback will be more likely to deliver a worse experience. It is important to stress that we do not claim in any way that this is the *optimal* measure of seller quality. The goal of our exercise is to show that our approach is feasible, and that it has broad implications for dealing with reputational externalities in platform markets. We revisit this issue in Section 7.

To illustrate our approach, consider two sellers: Seller A, who had 120 transactions, and seller B who had 150, yet both received one negative feedback and 99 positive feedbacks. Both have a PP measure of $\frac{99}{99+1} = 99\%$ and both have a score of $99 - 1 = 98$. Seller A, however, had only 20 silent transactions while seller B had 50 silent transactions. We define "effective" PP (EPP) as the number of positive feedback divided by total transactions, in which case seller A has an EPP of 82.5% while seller B has an EPP of only 66% and is a worse seller on average. Importantly, eBay does not display the total number of transactions a seller has completed and buyers cannot therefore back-out a seller's EPP score.

Figure 5 displays a histogram of EPP scores at the seller level from our dataset (as described below). The mean of this distribution is .64 and the median is .67. Importantly, unlike percent positive, there is substantial spread in the distribution.[12]

To verify that EPP contains information about buyers' experiences, we define a bad buyer experience (BBE) as one in which the buyer either left negative feedback, opened a dispute with eBay, or left low stars on the detailed seller ratings. In our data, which is described in detail in the next section, 3.39 percent of transactions resulted in BBEs. We run a probit regression of BBE on seller quality scores (EPP, PP, Feedback score) and controls for price, category, and purchase type (auction or fixed price). As table A-1 in the Appendix shows, the coefficient of interest on EPP is negative and highly significant, indicating that transacting with higher EPP sellers indeed decreases the probability that a BBE will occur, consistent with EPP being a measure of seller quality. We also show in the Appendix that EPP has as

[12]We also note that although PP and EPP are correlated, they are not overly so. A simple correlation coefficient between the two across sellers is 0.3.

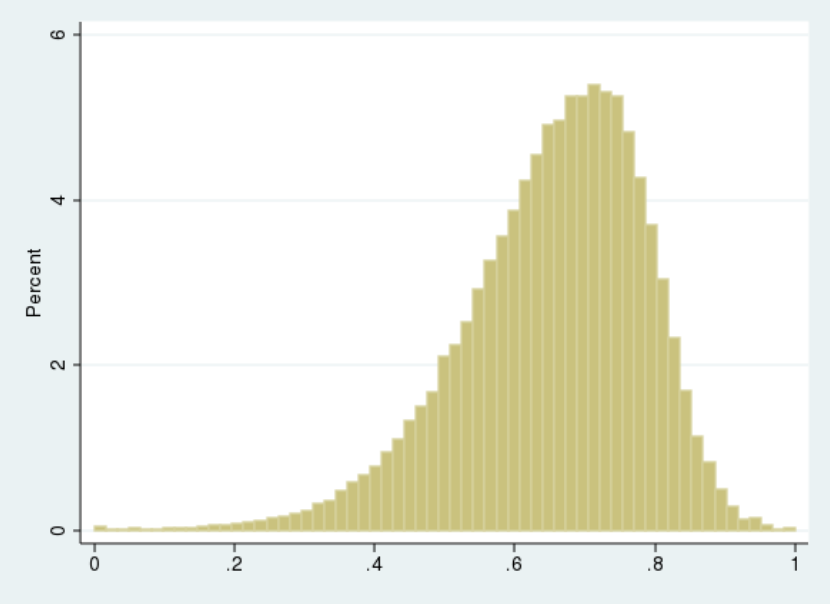

Figure 5: Histogram of Sellers' Effective Percent Positive Scores

much information as do the observable measures of reputation (PP and score), and that the information provided by EPP adds explanatory power above these two observable measures.

Even though EPP is unobservable, perhaps buyers observe signals that are correlated with EPP, questioning its exogeneity. The analysis of the controlled experiment in Section 6 should alleviate any such concerns. We also demonstrate in the Appendix that buyer experience is not correlated with EPP, further suggesting that EPP is exogenous to a buyer's choice.

4 Data

We selected users who signed up for a new eBay account on the U.S. site in 2011 and purchased an item within 30 days of sign-up. Because using all new buyers results in data that is too large for analysis, we used a 10% random sample and analyzed the behavior of a cohort of 935,326 buyers.[13] For each buyer we tracked all transactions from their sign-up until May 31, 2014, which resulted in 15,384,439 observations. Each observation contains rich

[13]Replicating the analysis for the 2008, 2009, and 2010 cohorts yields very similar results.

transaction-information including, but not limited to price, item category, title, the seller, whether it was an auction or fixed price, and quantity purchased.[14]

There were a total of 1,854,813 sellers associated with these transactions. We collected information on each of these sellers at the transaction level such as the feedback score, percent positive, and number of transactions the seller had in the past. We constructed a seller EPP score at each separate transaction by looking back at all of the seller's transactions (capped in January of 2005, the earliest data stored) up to the point right before the transaction. This generated a complete snapshot of the information structure at the point when the buyer was making his decision and, as such, we did not include the focal transaction in the measure. Recall that the buyers cannot observe or infer the EPP measure.

Figure 6 is a histogram of the total number of transactions by an individual buyer over the course of his tenure, with coarser bins for numbers of 25 and higher. A large percentage of eBay buyers made very few purchases over their life-cycle – a full 38% of new buyers purchased only once, with an additional 14% who purchased only twice. On the other end of the spectrum, there is an extremely large right tail. While the median number of transactions is 2, the mean is 16, the 95th percentile is 65, and the max is 19,359.

5 Reputational Externalities and Buyer Behavior

We begin by distinguishing between two scenarios. In the first, buyers use eBay merely to connect with certain sellers, and no externalities across sellers are present. In the second, buyers consider eBay as a provider of quality services, in which case they infer the quality of the platform from individual transaction and an externality across sellers exists.

Table 1 shows a cross tabulation by buyer of how the total number of transactions relates to the total number of sellers that a buyer interacted with. For example, of the 38,149 buyers who completed 20-29 transactions during our sample period, 23,367 bought from between 20

[14]Not all 15 million transactions were used in every regression because some transactions did not record all of the information we wished to include. For example, 2,127,108 of transactions do not contain item condition (new vs. used), which happens when a seller does not enter this data. We therefore excluded transactions for which we could not include the full set of covariates.

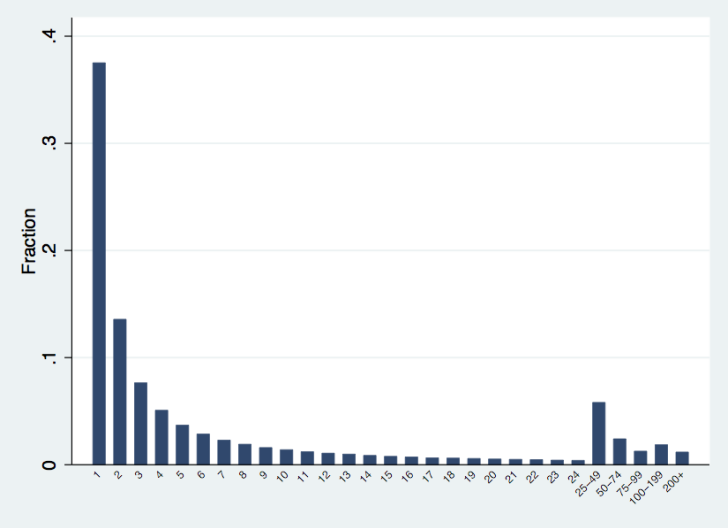

Figure 6: Histogram of Total Transactions by Buyer

and 29 different sellers while only 116 of them bought all their transactions from a single seller.[15] This shows that buyers tend to deal with large numbers of sellers and therefore suggests that externalities may indeed exist.

Table 1: Total Transactions by Total Number of Sellers for Each Buyer

		Total Number of Sellers						
		00-01	02-05	06-09	10-19	20-29	30-49	Total
Total Transactions	00-01	350,881	0	0	0	0	0	350,881
	02-05	27,603	253,032	0	0	0	0	280,635
	06-09	1,206	19,374	60,590	0	0	0	81,170
	10-19	492	2,802	15,959	64,112	0	0	83,365
	20-29	116	386	767	13,513	23,367	0	38,149
	30-49	67	207	273	1,810	11,685	24,106	38,148
	Total	380,365	275,801	77,589	79,435	35,052	24,106	872,348

To explore the extent of seller externalities, we ask how a transaction's quality affects the probability that a buyer will return to transact again with *any* seller, and compare this to the probability that he returns to transact with the same seller. If reputational externalities are

[15]The table uses 872,348 instead of 935,326 buyers because we included buyers with no more than 49 transactions for compactness.

present then a higher quality transaction will benefit the platform more than the individual seller, creating a wedge between the individual seller incentives and total welfare.

Our econometric specifications include regressions of the following form,

$$y_{ijt+1} = \alpha_0 + \alpha_1 EPP_{jt} + \beta \cdot \bar{b}_{it} + \gamma \cdot \bar{s}_{jt} + \delta \cdot \bar{d}_t + \varepsilon_{ijt}, \tag{1}$$

where y_{ijt+1} is an indicator equal to 1 if buyer i bought again on eBay after concluding transaction t with seller j, EPP_{jt} is seller j's EPP at transaction t, $\bar{b}_{it}$ is a vector of buyer characteristics (e.g., how many transactions they completed), $\bar{s}_{jt}$ is a vector of seller characteristics (e.g., reputation score and PP), and $\bar{d}_t$ is a vector of transaction characteristics.[16]

Because buyers do not observe EPP and do not act on it, we think of it as an exogenous seller quality shock. The higher the EPP, the more likely it is that the transaction goes well and therefore the more likely the buyer is to return and purchase on eBay – consistent with our conceptual framework outlined in Section 2.

Table 2 is our baseline regression table. Column 1 reports the results of an OLS regression where EPP enters linearly. Columns 2 through 4 divide EPP into its quartile breaks, allowing for non-parametric flexibility. Column 2 presents OLS estimates, column 3 presents probit estimates and column 4 shows the average marginal effects from the probit regression. Standard errors are clustered at the buyer level and the controls (transaction type, a buyer's past experience, item condition, the seller's experience level – how many transactions they've completed in the past, and the category of the item) are not reported for space considerations. We have run a further suite of regressions including separating out used vs. new items, including buyer fixed effects, and including seller fixed effects.[17]

The coefficient estimates are stable across the specifications and confirm the expected relationship. Seller quality, measured by EPP, is highly statistically and economically significant, indicating that when buyers purchase from a higher quality seller, they purchase

[16]The left hand side variable in these regressions is an indicator that a buyer purchases again on eBay within 180 days of the focal transaction. We have run this for other lengths of time including 60 days and whether a buyer ever returns to eBay. The results are qualitatively the same.

[17]These regressions provide further evidence for the exogeneity of EPP by relying on differing identification assumptions. Results are reported in the appendix.

Table 2: Baseline EPP Regressions

	OLS	OLS	Probit	Marginal Effects
EPP	0.139***			
	0.00112			
EPP Dummy (excluded: $0 < .517$)				
$\geq .517 < .592$		0.0192***	0.148***	0.0199***
		0.000253	0.00167	0.000225
$\geq .592 < .5668$		0.0289***	0.225***	0.0292***
		0.000285	0.00184	0.000239
$\geq .668$		0.0399***	0.309***	0.0385***
		0.000317	0.00211	0.000258
Seller Feedback Score	-8.86e-10	-1.52e-09	-0.000000113***	-1.39e-08***
	1.57e-09	1.55e-09	1.18e-08	1.45e-09
Percent Positive Dummy (excluded: $0 < .994$)				
$\geq .994 < 1$	-0.00931***	-0.00897***	-0.0522***	-0.00640***
	0.000210	0.000210	0.00149	0.000182
$= 1$	-0.0125***	-0.0102***	-0.0699***	-0.00864***
	0.000300	0.000295	0.00227	0.000285
Item Price	-0.000313***	-0.000316***	-0.00168***	-0.000207***
	0.00000382	0.00000381	0.0000131	0.00000162
Seller Standards Dummy (excluded: Below Standard)				
Standard	-0.00831***	-0.00840***	-0.0905***	-0.0108***
	0.000474	0.000474	0.00384	0.000449
Above Standard	-0.00788***	-0.00763***	-0.0672***	-0.00792***
	0.000412	0.000412	0.00338	0.000386
ETRS	-0.0118***	-0.0115***	-0.0898***	-0.0107***
	0.000425	0.000425	0.00340	0.000390
Constant	0.445***	0.506***	0.0714***	
	0.00104	0.000828	0.00495	
N	12,824,846	12,820,329	12,820,317	12,820,317

Controls for buyer number of transactions up to the focal transaction, new vs. used, auction vs. fixed price, product category, and number of seller transactions, are in the regression but not reported for brevity. See the appendix for robustness. Standard errors are clustered at individual level

more. Moving from the lowest quartile EPP seller to the highest quartile, increases the probability that a buyer purchases again within 180 days by around 4 percentage points.

Interestingly, the observable seller reputation measures are either unstable or mostly negative. We interpret this as selection: Once we control for seller quality using EPP, buyers self select into different sellers based on how they interpret quality or on their intentions. Perhaps, someone who is weary about eBay and does not plan to return may only choose to buy from 100% PP sellers, whereas someone who knows eBay and is planning on sticking around may be more willing to take a risk with a lower PP seller, causing the negative coefficient on PP=100%.

Table 3 repeats the regression but now y_{ijt+1} indicates if buyer i bought again *from seller* j after purchasing transaction t from seller j. The coefficients are qualitatively similar but quantitatively much smaller than they are in Table 2, indicating that the experience a buyer has on eBay is a lot more likely to influence whether he returns to the site, rather than to the same seller, which is a very unlikely event. This establishes that the extent of seller reputational externalities is significant.

The analysis above helps distinguish between two important reasons that a buyer may choose not to return to purchase. The first is selection: People come to eBay looking for a specific item, purchase that item and then have no need to return. The second is that buyers initially have limited knowledge about the platform and update beliefs over time, causing any seller's quality to influence their decision to come back to the platform.

The dynamic Bayesian updating framework described in Section 2 implies that transaction quality should matter less as a buyer completes more transactions. Every experience will help a buyer learn about his idiosyncratic match value with the site, as well as get a draw from the seller quality distribution on the site. Hence, a buyer with more experience is more likely, on average, to return to the site *both* because of selection (people who don't like the site have left already) and because a bad quality draw later will have less impact on his beliefs about quality because of a tighter posterior belief.

Table 3: Likelihood of Returning to the Same Seller

	OLS	OLS	Probit	Marginal Effects
EPP	0.0718***			
	0.00794			
EPP Dummy (excluded: $0 < .517$)				
$\geq .517 < .592$		0.00477**	0.0173**	0.00471**
		0.00154	0.00552	0.00150
$\geq .592 < .5668$		0.0212***	0.0664***	0.0183***
		0.00178	0.00631	0.00173
$\geq .668$		0.0199***	0.0740***	0.0205***
		0.00221	0.00766	0.00212
Seller Feedback Score	-0.000000386***	-0.000000385***	-0.00000119***	-0.000000328***
	2.15e-08	2.13e-08	5.78e-08	1.59e-08
Percent Positive Dummy (excluded: $0 < .994$)				
$\geq .994 < 1$	0.0329***	0.0320***	0.0950***	0.0268***
	0.00140	0.00140	0.00477	0.00134
$= 1$	-0.0359***	-0.0353***	-0.145***	-0.0379***
	0.00166	0.00162	0.00646	0.00165
Item Price	-0.000325***	-0.000326***	-0.00162***	-0.000448***
	0.0000151	0.0000151	0.0000788	0.0000217
Seller Standards Dummy (excluded: Below Standard)				
Standard	-0.0908***	-0.0908***	-0.384***	-0.0983***
	0.00232	0.00232	0.00832	0.00222
Above Standard	-0.00562**	-0.00534**	-0.0228***	-0.00653***
	0.00192	0.00192	0.00658	0.00190
ETRS	-0.00536*	-0.00512*	-0.0161*	-0.00463*
	0.00209	0.00210	0.00717	0.00207
Constant	0.138***	0.169***	-1.095***	
	0.00693	0.00490	0.0164	
N	11,883,455	11,879,306	11,879,303	11,879,303

Controls for buyer number of transactions up to the focal transaction, new vs. used, auction vs. fixed price, product category, and number of seller transactions, are in the regression but not reported for brevity. See the appendix for robustness. Standard errors are clustered at individual level

In the Appendix we show evidence consistent with EPP having a weaker effect on the choices of more experienced buyers. While we cannot separate the effects of selection from the effects of Bayesian learning, it is reassuring that the effect of EPP on buyers is consistent with our simple framework.

6 Using Search to Internalize Seller Quality

This section reports results from an experiment in which our measure of seller quality, EPP, was incorporated into eBay's search algorithm to promote higher EPP sellers at higher positions on the search results page (holding everything else constant) for a randomly selected group of users. The experiment allows us to do three things: (1) Answer any lingering doubts about the exogeneity of EPP as an unobserved measure of seller quality; (2) Explore the extent to which consumers respond to search ranking schemes, and hence how effective changes in them might be for platforms wishing to internalize seller quality externalities; and (3) Quantify the downside of using search rankings – the extent to which consumers do not purchase because they are unable to find the product they are looking for.

Two issues are worth noting. First, for this strategy to work, buyers must be more likely to select an item higher up on the search results page, implying some sort of search costs. The literature on search costs has demonstrated correlation between ranking and purchase (or click-through) behavior (Ghose et al., 2013). Second, promoting seller quality may come at the expense of providing fewer relevant items. Thus, as mentioned earlier, the third objective of the experiment is to estimate the trade off inherent in manipulating search results to prioritize better quality sellers. The long-term benefit from buyers interacting with better quality sellers and returning to the site must be weighed against the short-term loss of buyers being less likely to purchase because they do not find what they want.

Figure 7 shows an example of the eBay search results page (SRP) returned to a user who searched for the query "pink bunny rabbit slippers." The SRP is a mix of different products that eBay's search algorithm matches to the buyer's query. In our treatment group, sellers with higher EPPs will be promoted and those with lower EPPs demoted, necessarily

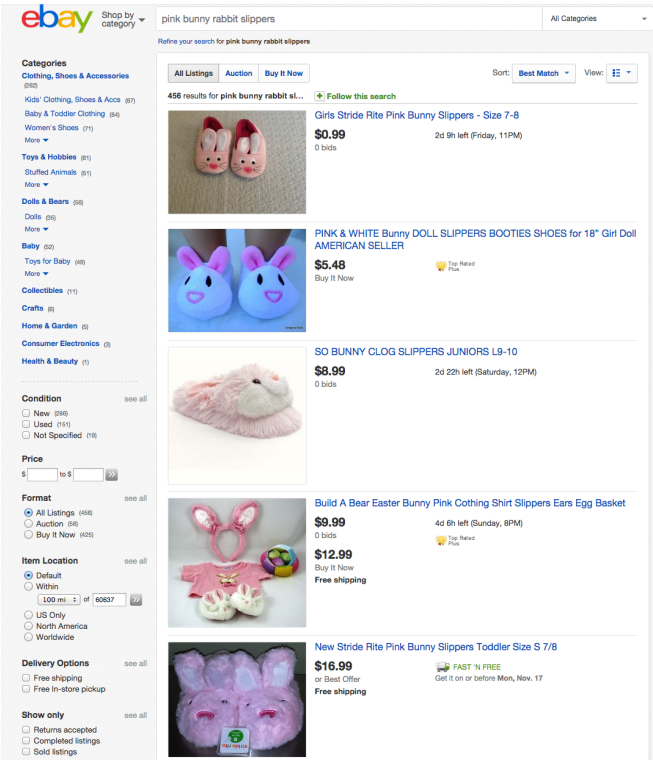

Figure 7: The eBay Search Results Page

leading to different products being ranked higher (except in the almost nonexistent case where all products returned for a given query are homogeneous). Thus, the experiment cleanly promoted higher quality sellers (according to our EPP measure) by changing which products were displayed at the top of the SRP.

The experiment ran from December 14th, 2011 though January 2, 2012 on 10% of eBay's U.S. site traffic—several million searches per day—that was placed into our experimental treatment and exposed to a ranking scheme that differed from the default site algorithm. Because of other site optimization considerations, we had limited control over the weighting that the EPP measure received, a point to which we will return below.

The ideal treatment would ensure that an individual user was always in the treatment group whenever he searched for a product on eBay. However, it is impossible to unambiguously link a site visit to a specific user either because the user visits the site from a computer or browser he has never signed in from before, or because he has deleted his cookies (the way

sites keep track of users between visits). This creates the potential for leakage between the treatment and control groups, where a user is sometimes exposed to search results in the treatment group and sometimes in the control group.

To understand the magnitude of this problem, it is necessary to understand how site visits are linked to users. eBay stores a Globally Unique Identifier (GUID) in a cookie on browsers that visit the site, allowing it to track whether the same browser visits the site again. An algorithm attempts to match GUIDs to user IDs (UIDs) by tracking whether that browser was used to sign into an eBay account at any time. Multiple GUIDs may be linked to the same UID if a user signs in from multiple browsers on the same computer or from multiple computers. Our experiment was run at the GUID level, meaning that 10% of active GUIDs were placed into the treatment group. A user therefore could be placed into the treatment group for one, but perhaps not all, of the GUIDs linked to his account. Fortunately, we can track this behavior and observe the number of searches that a user made within the treatment and control groups. We limit our analysis to users that come from only one GUID, i.e., all of their searches are either in the control group or treatment group.

Table 4: User level summary statistics

	Treatment	Control
Number of Users	1,258,455	11,486,810
Total Searches	46,015,313	417,284,312
Avg. Number of Searches	36.565	36.327
	(136.979)	(118.782)
Avg. Number of Sessions	3.617	3.616
	(6.260)	(6.247)
Avg. Number of Purchases (during experiment)	0.554	0.551
	(1.826)	(1.849)
Avg. Number of Past Trans	62.101	62.002
	(183.216)	(179.763)

Table 4 displays basic summary statistics for users both before and during the experiment, confirming that the assignment was indeed random (none of the differences are statistically significant from each other). It's also worth noting the large heterogeneous variance across users. On average a user in the treatment (control) group performed 36.6 (36.3) searches during the two week experiment, but with a huge standard deviation of 137 (118.8).

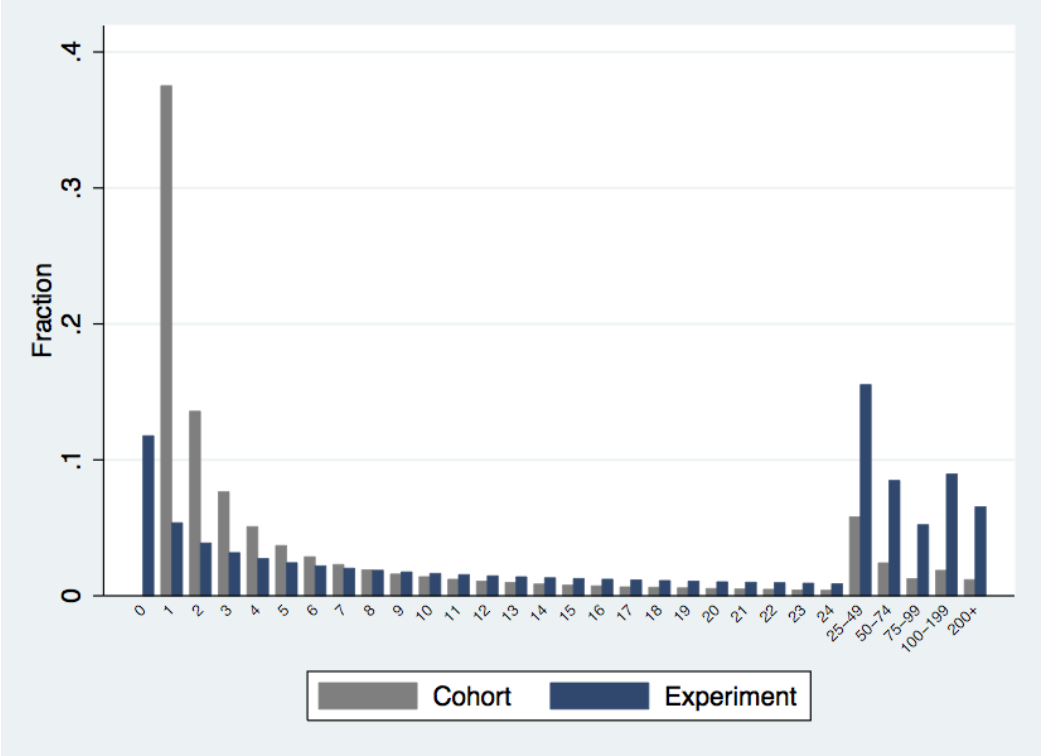

Figure 8: Comparison of purchases between cohort and experiment samples

The distribution of users is different than in our cohort study described in the previous section because the cohort study was carefully structured to follow a user through his tenure on eBay, while the experiment contains users who searched for a product during the two week experimental period. Figure 8 plots the distribution of past user transactions across users in the experiment (extending the information contained in the last row of table 4) and compares it to the cohort distribution. Users in our experiment are more active users with a median of 23 purchases, relative to a median of 2 in the cohort sample, which is expected of a random sample of users who visit eBay within any two week time period. We also note that our experiment took place at the end of 2011, relatively at the beginning of the cohort transactions. For these reasons, the results from the experiment may not perfectly match the results from the cohort study.

We collected data on all searches performed during the experimental period (including the query and the items that were displayed), whether or not the GUID was in the treatment

or control group, all other user behavior (clicks on products, etc.), and purchases both during the experimental period and after the experimental period. We test whether a buyer is more likely to purchase in the future if randomly assigned to the treatment group, conditional on purchasing in the experimental period.

Because an individual search returns several items (typically 50) on a single page, each associated with a seller (and EPP score), it is instructive to collapse a whole search into a single measure of the quality of sellers in the search. We define "Discounted Search EPP" (DSEPP) where each item is weighted by its position in the search results. Specifically, we weight the EPP score of each item displayed to the user by the inverse of the item's position on the search results page. This reflects the prevailing belief that items ranked higher up on the page are more visible, and hence play a larger role in the user's decision process.

Figure 9 shows a kernel density plot of the DSEPP scores for all searches in the treatment and control groups. The mean in the control group is 60.13% and of the treatment group is 61.85%, and the distributions are statistically different from each other. Average EPP for purchases in the control group was 61.57% compared with 62.27% in the treatment group. The correlation between search EPP and purchase EPP was 0.68, meaning higher EPP in search results translated into buyers transacting with higher EPP sellers.

Conceptually, we view the true treatment effect as coming from being matched to, and purchasing from, a higher quality seller. Because our experiment randomized at the search and not at the purchase level, our experiment can be viewed as an intent to treat design with potential selection into who was actually treated (i.e., who actually purchased from a higher quality seller). One might be concerned that the group that purchases during the experiment in the treatment group (and hence is truly treated) is somehow different than those that purchase during the experiment and were in the control group. For example, if increasing EPP came at the expense of search relevance, those that do purchase in the treatment group may be differentially selected to be loyal eBay users, and are thus more likely to come back in the future regardless of the true treatment effect.

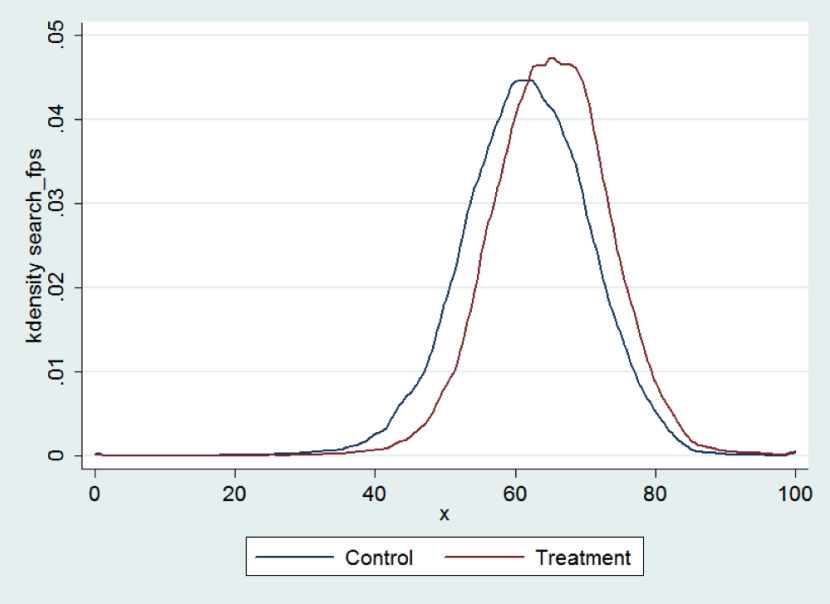

Figure 9: DSEPP Scores between Treatment and Control Groups

We thus analyzed the experimental data in three steps: First, we compared future purchase behavior between the treatment and control groups regardless of whether or not they purchased during the experimental period. This analysis solely exploits the experimental randomization and we see it as a lower bound of the true treatment effect size (because it mixes those who purchased, and thus could be considered truly treated, and those that did not).[18] Second, we analyze the sequential behavior between search and purchase, calculating whether consumers in the treatment group are disproportionately less likely to purchase, conditional on search. We find no evidence that this is the case and hence the argument for selection into true treatment is limited. Third, we analyze the probability of return conditional on purchase during the experimental period, controlling for observables, including characteristics of the buyer, seller, and the transaction, and using the experiment as an

[18]For example, consider the selection story above. If the set of individuals who chose not to purchase in the treatment group, but would have purchased in the control group, do not have a higher propensity of purchasing again from eBay *because* they didn't purchase during the experiment, then our intent to treat estimates will be a lower bound. We believe this to be the case because these users would be dissatisfied with not finding what they were looking for during the search attempt, making them less likely to return to eBay. Furthermore, because users were not aware that the experiment was occurring, they wouldn't know to alter their behavior to avoid the experiment (by, say, returning when they knew the experiment was going to be over and search results were going to be returned to normal).

instrumental variable. We view the first of these analyses as an intent to treat estimate and the third of these as the true treatment effect on the treated.

6.1 Intent to treat estimate

Table 5: Two-sample test of proportions

Group	Obs	Mean	Std. Err.	95% Conf.	Interval
Control	11,486,810	.6155062	.0001435	.6152249	.6157875
Treatment	1,258,455	.6185275	.000433	.6176788	.6193762
diff		**.0030213**	.0004562	.0021272	.0039153
diff	prop(1) - prop(0)				**z = 6.6151**

Table 5 shows the probabilities of return divided into the treatment and control samples. The difference is a statistically significant 0.3 percentage points. Magnitudes should be judged in the context of the size of the shift in DSEPP that the experiment created. We calculated the raw marginal change in probability of return normalized by the shift in DSEPP as:

$$\frac{\Delta Pr\{\text{return}\}}{\Delta\text{DSEPP}} = \frac{(0.6185275 - 0.6155062)}{(0.6227 - 0.6157)} = 0.43$$

This estimate is about three times higher than that of the cohort analysis described in column 1 of Table 2 (0.43 instead of 0.139 in the cohort analysis). As we show below, controlling for observables brings the experimental estimate a lot closer to that of the cohort analysis.

As corroboration of the raw intent to treat estimate, we break apart buyers into quartiles based on their tenure on eBay (up to the beginning of the experiment). We compare the difference in probability of return by treatment versus control in the top quartile of users (those with 60 or more transactions on eBay) to that of those in the bottom quartile (those with 4 or fewer transactions). Bayesian learning implies that those in the bottom quartile – with few transaction on eBay – should be more affected by being in the treatment group compared to those in the top quartile. Tables 6 and 7 show the difference for those in the bottom and top quartiles respectively.

Table 6: Bottom quartile of buyers by past transactions

Group	Obs	Mean	Std. Err.	95% Conf.	Interval
Control	3,095,706	.2935343	.0002588	.2930271	.2940416
Treatment	335,752	.296028	.0007878	.2944839	.2975721
diff		**.0024937**	.0008293	.0008684	.004119
diff	prop(1) - prop(0)				**z = 3.0131**

Table 7: Top quartile of buyers by past transactions

Group	Obs	Mean	Std. Err.	95% Conf.	Interval
Control	2,904,679	.875757	.0001935	.8753777	.8761363
Treatment	319,702	.8760627	.0005828	.8749205	.8772049
diff		**.0003057**	.0006141	-.0008979	.0015092
diff	prop(1) - prop(0)				**z = 0.4974**

The tables show that those in the bottom quartile are statistically significantly affected by the treatment while those in the top quartile are not.[19] To test whether these differences are statistically significant, Table 8 shows the results of a regression of the probability of return on a dummy for being in the treatment group, a dummy for being in the top quartile (relative to the bottom quartile – those in the middle quartiles are excluded from the regression) and the interaction of the two. The regression shows that the interaction is significant at the 5% level (a t-stat of 2.10), consistent with the Bayesian learning framework that we use.

Next, we replace the two-sample test of proportions with a regression form, controlling for characteristics of the users in the treatment versus the control group. In particular, we control for the pre-experiment purchase behavior of users and their behavior during the

[19]It is a bit curious that the intent to treat estimate is actually smaller for the bottom quartile than it is for the overall population – table 5 above. We note that this difference is not statistically significant (t = 0.96) and that this is mathematically possible because the split is based on a variable that is not the computation in the t-test.

Table 8: Intent to treat estimates by quartile

LHS: Prob of return	b/se
Treatment dummy	
Excluded: Control	
Treatment	0.00249***
	0.000726
Top quartile dummy	
Excluded: Bottom quartile	
Top quartile	0.582***
	0.000326
Interaction dummy	
Top quartile * treatment	-0.00219*
	0.00104
Constant	0.294***
	0.000227
N	6,655,839

experiment (number of searches and number of sessions). Table 9 displays these results. The estimates shrink when controlling for observables but remain statistically significant.[20]

Using the OLS estimate of 0.001 from column 1 of Table 9 instead of the estimate of 0.003 from Table 5, and recalculating the raw marginal change in probability of return normalized by the shift in $DSEPP$ gives a value of 0.157, a lot closer to the 0.139 from Table 2.

6.2 Sequential Behavior from Search to Purchase

Next, we examined the impact that changing search results had on purchase behavior during the experimental period. It is possible that changing search results in the treatment group might reduce relevance (relative to the control group which was optimized to maximize the short term probability of purchase) and thus lead to a lower probability of purchase conditional on search. We found no evidence that this occurred and the probability of purchase did not decrease in the treatment group.

[20]This is not necessarily surprising. Small imbalances in sample characteristics can cause differences in outcomes even with large samples such as we have here. In this case, past purchase behavior on eBay, which was slightly higher in the treatment relative to the control group, might be causing these differences.

Table 9: Probability of return in 180 days

	OLS b/se	Probit b/se	Marginal Effects b/se
Treatment Dummy	0.00106**	0.00372**	0.00112**
	0.000398	0.00131	0.000396
Past Transactions excluded: 0			
1-4	0.153***	0.471***	0.142***
	0.000461	0.00150	0.000448
5-22	0.357***	0.995***	0.300***
	0.000417	0.00137	0.000386
23-71	0.533***	1.502***	0.453***
	0.000423	0.00142	0.000368
72-172	0.611***	1.836***	0.554***
	0.000480	0.00172	0.000448
173-287	0.617***	1.908***	0.575***
	0.000688	0.00269	0.000767
Number of Searches excluded: 1			
2-3	0.00537***	0.0136***	0.00410***
	0.000472	0.00149	0.000449
4-12	0.0269***	0.0735***	0.0222***
	0.000440	0.00139	0.000419
12-41	0.0553***	0.158***	0.0476***
	0.000490	0.00156	0.000469
42-119	0.0814***	0.257***	0.0774***
	0.000616	0.00203	0.000612
Constant	0.144***	-1.035***	
	0.000479	0.00159	

Figure 10 explores the difference between the treatment and control groups in the probability that a session that contains a search in either the treatment or control groups ends in a purchase. The panel on the left side plots the raw probabilities by treatment and control groups – the red line for the control group, and the blue line for those in the treatment group. As is easily seen, the lines are right on top of each other. To further investigate this, the panel on the right plots the difference between the treatment and the control group with dashed lines for the 95% confidence interval. The difference in conversion probability is precisely estimated at 0. Consumers are purchasing from different sellers (and potentially different products), but the probability of purchase was not affected. Hence, we are not concerned that the treatment effects (conditional on purchase) are a result of selection into who purchases.

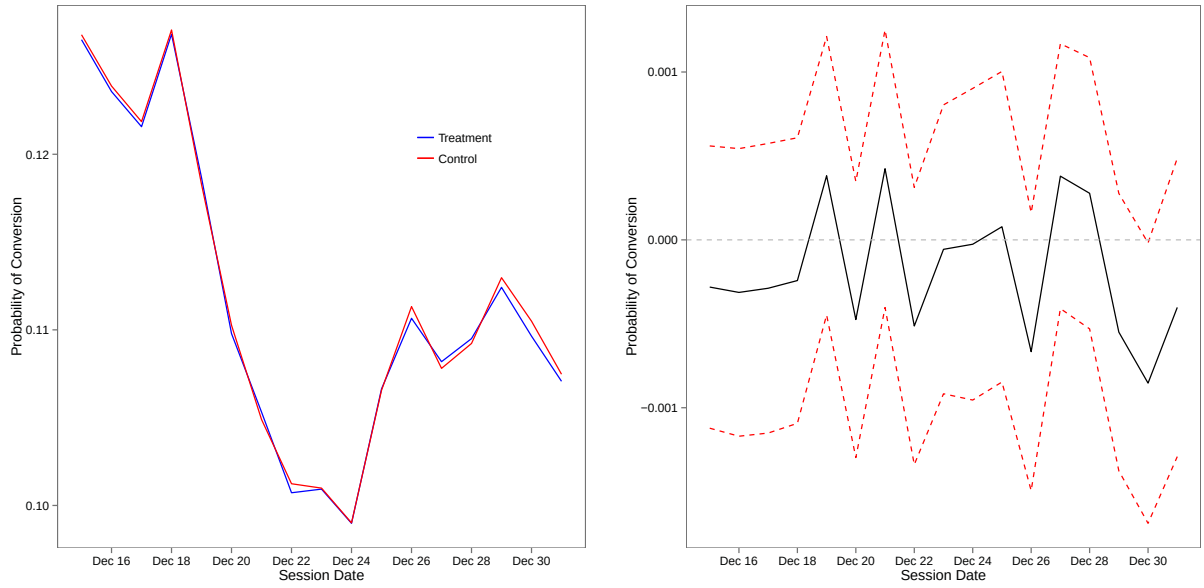

Figure 10: Probability of Purchase between Treatment and Control Groups During the Experiment

6.3 Treatment Effect on the Treated

Next, we examined the effect of an increase in EPP (via treatment) on the probability of return, *conditional* on a purchase during the experiment. Out of the experiment's 12,745,265 users, 9,120,925 (72%) did not purchase during the experimental time period. Here we focus on the 3,624,340 (28%) of users who purchased one or more times. This gives us a total of 5,502,532 transactions for which we had a complete set of covariates in the eBay data warehouse. The analysis that follows includes only these transactions. Conditional on purchase, we ran a series of regressions to measure the effect of EPP on the probability of return, leveraging the clean experimental variation.

Column 1 in Table 10 regresses a return indicator (within 180 days) on a dummy for whether the transaction came from a search in the treatment group or the control group, controlling for the same characteristics as in the cohort study (Table 2). Column 2 repeats the same regression as in the cohort study, where EPP enters in linearly. Columns 3 and 4 use the experimental variation as an instrument for EPP under the assumption that searching in the experiment exogenously led to a higher purchase EPP, which is consistent with the data.

Table 10: Probability of return in 180 days

	ols b/se	ols b/se	firststage b/se	ivresults b/se
EPP		0.261***		0.246***
		0.00174		0.0985
Treatment Dummy	0.00137**		0.00557***	
	0.000550		0.000134	
Seller Feedback Score	8.94e-09***	5.64e-09***	1.27e-08***	5.83e-09***
	6.07e-10	6.07e-10	1.48e-10	1.39e-09
Percent Positive Dummy excluded: $0 < .994$				
$\geq .994 < 1$	0.0145***	-0.00760***	0.0847***	-0.00631
	0.000403	0.000429	0.0000984	0.00835
$= 1$	0.0203***	-0.00740***	0.106***	-0.00579
	0.000563	0.000592	0.000137	0.0105
Item Price	-0.0000662***	-0.0000624***	-0.0000144***	-0.0000626***
	0.000000943	0.000000941	0.000000230	0.00000170
Seller Standards Dummy excluded: Below Standard				
Standard	-0.0420***	-0.0366***	-0.0208***	-0.0369***
	0.00116	0.00116	0.000284	0.00236
Above Stand	-0.0208***	-0.0197***	-0.00433***	-0.0198***
	0.00106	0.00105	0.000258	0.00114
ETRS	-0.0383***	-0.0339***	-0.0166***	-0.0342***
	0.00105	0.00105	0.000256	0.00195
Constant	0.782***	0.634***	0.566***	0.643***
	0.00108	0.00146	0.000265	0.0558
N	5,502,532	5,503,316	5,502,532	5,502,532

Controls for buyer number of transactions up to the focal transaction, new vs. used, auction vs. fixed price, product category, and number of seller transactions, are in the regression but not reported for brevity. See the appendix for robustness. Standard errors are clustered at individual level

Column 3 is the first stage of that IV regression, EPP regressed on a dummy for whether or not the search was in the treatment group, controlling for all of the other predictors in the regression. The effect is strong and highly significant (a t-statistic of 41.52). The coefficient on EPP in the IV regression (column 4) is close to the OLS estimate in column 2, indicating that our exogeneity assumption in the cohort analysis is well-founded. Note the standard error increases substantially relative to the OLS regression, indicating the experimental instrument is not weak (in the sense of a low first stage T-stat) but is not extremely highly correlated with EPP (the coefficient size is .5 percentage points).

Together with the reduced form results of Section 5, the experiment demonstrates three important facts. First, a transaction's quality is an important component of an individual's likelihood to return to the platform, over and above his propensity to return to an individual seller. Second, platforms can use an intermediate screening mechanism – search result rankings – to guide buyers to better quality sellers, alleviating some of the externality issues associated with platform transaction quality. We note that search ranking may be one of many levers that platforms might use. Third, search ranking causally affects buyer purchase decisions. We show this by varying the search ranking of the same search in treatment versus control groups which gets around the traditional problem of search ranking endogeneity.

7 Discussion

A well-functioning reputation mechanism allows buyers to correctly infer the likelihood of a transaction going well without having past experience with any particular seller. The extent to which reputation mechanisms work depends on three important assumptions. First, that sellers indeed internalize the effect of their actions on future outcomes. Second, that the public information correctly mirrors the quality of past transactions. Third, that buyers correctly interpret reputation information. If any of these assumptions fail then buyers will inaccurately infer individual seller and aggregate platform quality.

We demonstrated that in practice, these assumptions are hard to satisfy in market platforms. First, there is a reputational externality across sellers, and second, reputation feedback can be—and in eBay's case is—biased. We studied the limits of reputation mechanisms in the face of these problems and their impacts on the marketplace. We then offered an implementable search prioritization strategy that online platforms can use to mitigate the adverse impacts of reputational externalities and biased feedback, and demonstrated its effectiveness through a field experiment on eBay's platform.

It is important to emphasize that EPP is but a small illustration, or "proof of concept," of the approach we are advocating for. Other sources of data can surely improve on the platform's ability to estimate seller quality. For example, another source of unobserved

information at eBay is obtained from the fact that it tracks email messages between buyers and sellers. Masterov et al. (2014) show that the content of these messages includes additional valuable information about the quality of the transaction above and beyond the observable information and EPP. It surely will be the case that different platforms will likely have different sources of internal data that can be used to estimate the quality of transactions. Hence, the form of the optimal estimator in different platform markets is a question of statistical fit and engineering, informed by economic theories of buyer and seller behavior.

We further advocate that online marketplaces use measures like EPP in more opaque ways that improve a buyer's experience indirectly through the marketplace's search rank algorithm, rather than display them directly to buyers, for two reasons. First, different buyers may interpret the same information in different ways. For one buyer a score of 88% might be satisfactory, while for another it is not, without having a clear understanding of how such a score translates into actual experiences. In theory, every rational expectations model of reputation has buyers being fully informed about the relationship between scores and outcomes, but in practice, and especially for less experienced buyers, such a mapping is unlikely to exist. Second, by making measures like EPP observable to buyers, sellers will most likely harass buyers who do not leave feedback in order to manipulate this new measure of seller quality. This would then cause a bias in EPP, and other measures of seller quality will need to be inferred from other parts of the data. This observation has another important practical implication. If, over time, sellers learn the ways in which the platform uses measures such as EPP to rank them, then they may engage in activities that reduce the informativeness of these measures. This suggests that platforms may have to continually search for better internal measures of seller quality, a consequence of a "cat and mouse" chase between the platform and its sellers.

One nice feature of our search prioritization strategy is that it differentially affects buyers with different search costs. We presented evidence of buyer learning about the platform. If search costs are correlated with experience on the platform, which we suspect they might be, then our intervention naturally separates out new buyers from more experienced ones as the

search costs of the latter ought to be lower. Specifically, if experienced buyers are more likely to search extensively (because they have lower search costs driven by more familiarity with eBay), then the intervention should affect them less. This is exactly the strategy a platform would like to implement because it exposes new buyers to better quality sellers.

We argued that factors such as reputational externalities across sellers play an important role in influencing market platforms, and hence delivering welfare to consumers. The theoretical literature on two-sided markets has, however, generally ignored this issue. The standard model that Rochet and Tirole (2006) propose for two-sided markets does not allow for externalities between agents on the same side of the market and concentrates on the binary decision of joining a platform or not. While these models may be appropriate for industries such as credit cards (the classic example used in many of these papers), the models cannot capture the complexity of relationships that exist on large marketplace platforms. In light of the growing importance of online platform markets in the economy, we advocate allocating some focus to models that can incorporate more complex market setups.

References

Bajari, P. and Hortaçsu, A. (2004). Economic insights from internet auctions. *Journal of Economic Literature*, 42(2):457–486.

Bar-Isaac, H. and Tadelis, S. (2008). Seller reputation. *Foundations and Trends® in Microeconomics*, 4(4):273–351.

Bolton, G., Greiner, B., and Ockenfels, A. (2013). Engineering trust: reciprocity in the production of reputation information. *Management Science*, 59(2):265–285.

Cabral, L. and Hortaçsu, A. (2010). The dynamics of seller reputation: Evidence from ebay*. *The Journal of Industrial Economics*, 58(1):54–78.

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

Dellarocas, C. and Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, 54(3):460–476.

Ghose, A., Ipeirotis, P. G., and Li, B. (2013). Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science, Forthcoming*.

Greif, A. (1989). Reputation and coalitions in medieval trade: Evidence on the maghribi traders. *The Journal of Economic History*, 4(4):857–882.

Hui, X.-A., Saeedi, M., Sundaresan, N., and Shen, Z. (2014). From lemon markets to managed markets: the evolution of ebay's reputation system. *Working paper, Ohio State University*.

Klein, T., Lambertz, C., and Stahl, K. (2013). Adverse selection and moral hazard in anonymous markets. *CEPR Working Paper*, (DP9501).

Luca, M. (2014). Reviews, reputation, and revenue: The case of yelp.com. *Working paper*.

Masterov, D., Tadelis, S., and Mayer, U. (2014). Canary in the e-commerce coal mine: Detecting and predicting poor experiences using post-transaction buyer-to-sellermessages. *Working paper*.

Mayzlin, D., Dover, Y., and Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8):2421–55.

Rochet, J.-C. and Tirole, J. (2006). Two-sided markets: A progress report. *The RAND Journal of Economics*, 37(3):645–667.

Tirole, J. (1996). A theory of collective reputations (with applications to the persistence of corruption and to firm quality). *Review of Economic Studies*, 63(1):1–22.

Appendix

1 Further Analyses of EPP

This section provides further evidence that EPP contains information related to transaction quality and that this quality is indeed not observable to buyers.

Table A-1 contains results from a regression of whether a transaction ended as a bad buyer experience (defined in the main body of the paper above) on EPP and a host of control variables. We urge caution in interpreting these results as the reporting of BBEs by users is highly selected. Buyers must take some action to alert eBay that the transaction went badly. This happens relatively rarely – around 3.5% of the time. Nevertheless, the variables move in expected directions. Transactions where sellers have higher PP, feedback scores, or are above standard or ETRS are less likely to end in BBEs. Importantly, the coefficient on EPP is strongly significant and negative indicating that BBEs are much less likely when transacting with high EPP sellers.

We have also run variations on the regression reported in A-1 to examine the additional variation that EPP adds in explaining BBEs over and above observable feedback measures. A linear probability regression that simply controls for characteristics of the transaction and seller without observable feedback measures or EPP yields an R-squared of .0049. A regression with observable feedback measures (PP and feedback score) has an R-squared of .0065; one with EPP but no observable feedback measures gives .0066; and one with both EPP and observable feedback measures gives .0074. We interpret these results as indicating that EPP provides as much predictive power in explaining BBEs as observable feedback measures and provides substantial predictive power on top of observable measures.

Table A-2 examines whether buyers who are more experienced (have completed more transactions on eBay) differentially select higher EPP sellers. If this were the case we might worry that buyers who are more knowledgeable about eBay, proxied by experience, are able to read between the lines and discover sellers of higher quality and hence with higher EPP scores. This would raise concerns about any causal interpretation of the EPP coefficients in the main

Table A-1: Probit of Bad Buyer Experiences and EPP

LHS: BBE Flag (0/1)	b/se
Seller Feedback Score	-6.55e-08***
	5.21e-09
Percent Positive Dummy	
excluded: $0 < .994$	
$\geq .994 < 1$	-0.146***
	0.00284
$= 1$	-0.182***
	0.00336
EPP	-0.695***
	0.0134
Item Price	0.00113***
	0.0000260
Seller Standards Dummy	
excluded: Below Standard	
Standard	0.0448***
	0.00588
Above Standard	-0.129***
	0.00551
ETRS	-0.278***
	0.00580
Constant	-1.278***
	0.0135
N	12,814,847

Regression includes controls for (coefficients not displayed) auction type (auction, fixed price), item category, item condition (new, used, refurbished), and buyer transaction number

regressions. Here we use EPP as the dependent variable, regressed against buyer experience level (transaction number) and controls for other characteristics of the seller and transaction. The coefficient on buyer transaction number is *negative*, indicating that, if anything, more experienced buyers select lower EPP sellers. We note that even if the coefficient is statistically significant, its economic magnitude is incredibly small. A buyer moving from an experience of 1 transaction to 1,000 transactions (well outside of the scale of most buyers in our sample) is correlated with a lower EPP score of only .004 percentage points. We interpret this as finding no evidence that buyers with more experience are able to differentially select higher (or lower) EPP sellers, providing evidence that EPP is unobservable to buyers.

Table A-2: Selection into EPP

LHS: EPP	b/se
Buyer Transaction Number	-0.00000405***
	4.28e-08
Seller Feedback Score	0.000000116***
	3.74e-10
Percent Positive Dummy	
excluded: $0 < .994$	
$\geq .994 < 1$	0.0711***
	0.0000619
$= 1$	0.0842***
	0.0000949
Seller Standards Dummy	
excluded: Below Standard	
Standard	-0.0184***
	0.000164
Above Standard	-0.00996***
	0.000144
ETRS	-0.0126***
	0.000145
Item Price	-0.0000959***
	0.000000626
Constant	0.634***
	0.000182
N	12,814,870

Regression includes controls for (coefficients not displayed) auction type (auction, fixed price), item category, item condition (new, used, refurbished), and the number of transactions a seller has completed up to the focal observation

2 Bayesian updating on platform quality

In this section we consider the interaction between EPP and a buyer's experience level on the platform. One implication of our bayesian updating story is that a buyer's probability of return will be less affected by transaction quality later in his tenure. We provide evidence consistent with this hypothesis although we note that our results are also consistent with a selection story.

Table A-3 cross tabulates the buyer's experience measured in the number of transaction against whether or not a buyer returns to purchase on eBay within 180 days. As a buyer

becomes more experienced, he is much more likely to return to eBay and purchase, consistent with either selection or learning about the idiosyncratic match with the platform.

Table A-3: Cross Tab: No. of Transaction with 180-day Return Probability

	No Return	Return	Total
01-05	912,616	1,788,930	2,701,546
	33.78%	66.22%	100.00
06-09	135,472	938,428	1,073,900
	12.61%	87.39%	100.00
10-19	132,613	1,650,773	1,783,386
	7.44%	92.56%	100.00
20-29	55,264	1,146,697	1,201,961
	4.60%	95.40%	100.00
30-49	50,728	1,557,494	1,608,222
	3.15%	96.85%	100.00
50-99	41,639	2,082,843	2,124,482
	1.96%	98.04%	100.00
100+	32,002	4,858,940	4,890,942
	0.65%	99.35%	100.00
Total	1,360,334	14,024,105	15,384,439
	8.84	91.16	100.00

Next, we consider the effect that a transaction's quality has on a buyer as he becomes more experienced. We modify the analysis to interact EPP with a buyer's experience, measured by the number of transactions that a buyer completed including the focal transaction. The specification is as follows,

$$y_{it+1} = \alpha_0 + \alpha_1 EPP_{jt} * Tr_{it} + \beta \cdot \bar{b}_{it} + \gamma \cdot \bar{s}_{jt} + \delta \cdot \bar{d}_t + \varepsilon_{ijt}$$

where Tr_{it} is a dummy variable that takes on the transaction number for an individual buyer. This measures the strength that EPP has on the probability of return separately for a buyer at each point in his tenure. We could have very similarly ran separate regressions for each buyer transaction number. This regression specification contains two types of variation: First, for a given buyer "type", transactions occur at different points in his tenure, i.e., someone who will eventually transact 25 times on the site, may react differently to EPP at different points along his tenure. Second, different buyer "types" figure into the regression differently.

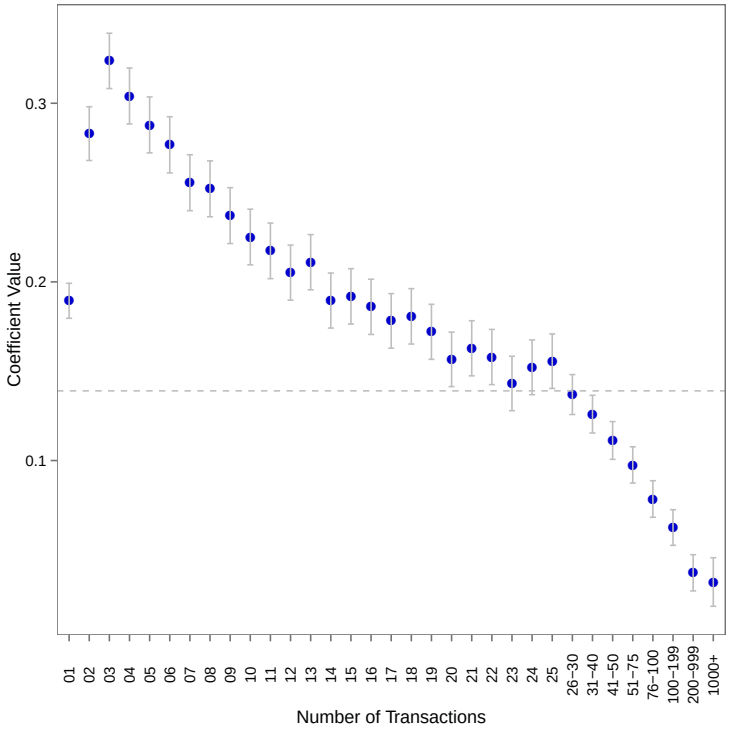

Figure A-1: The differential effect of EPP across a buyer's tenure on eBay

For instance, the coefficient on EPP interacted with $TR_{it} = 25$ only contains buyers that have stayed on the site to conduct at least 25 different transactions.

Figure A-1 plots the marginal effects of the coefficients and standard errors from this specification. For comparison purposes, the dashed line plots the average effect of EPP from a regression without the interactions, i.e., where EPP is constrained to be the same for all users. EPP has a statistically significant effect on the probability of return as far out as a buyer's $1,000^{th}$ transaction on eBay and the magnitudes vary dramatically, ranging from a high of 0.33 for the third transaction to a low of 0.02 for transactions over $1,000$. The coefficients confirm that EPP matters much more for transactions that occur early on in a buyer's tenure, consistent with our Bayesian-learning framework. Interestingly, the effect is not monotonic for the first and second transactions. We interpret this as people who intend to purchases once or twice and not to return regardless of the transaction's quality. We believe that these are people who need a very specific item that only eBay carries but otherwise would avoid

the site. EPP still matters for these early transactions, it just matters less because of the mix of these one-time users with the users who consider eBay as a long-term destination and are therefore updating on the overall platform quality.

3 Robustness

In this section we explore a series of robustness checks and additional specifications. For all of these regressions, the dependent variable is a binary indicator of whether a buyer purchases again within 180 days. For compactness we do not report the results for 60 days and whether the buyer ever returns to eBay, but results are very similar for those variables.

Table A-4 breaks out the main regression for new (column 1) versus used (column 2) products. The EPP coefficient is larger for new products than for used ones. To us, this illustrates a central challenge in this line of research, namely buyer expectations and heterogeneity. One might initially think that EPP would matter more for used rather than new transactions because the variance in transaction quality could be higher. Offsetting this, however, is the fact that different buyer types purchase new items vs. used items and that buyers have different expectations over these transactions. A bad experience with a used good might prompt a buyer to say "Oh, well, I knew I was taking a chance on a used good" whereas a bad transaction on a new good might prompt buyer exit.

Table A-5 displays specifications with seller fixed effects (column 2) and buyer fixed effects (column 3). Note that both because of computational issues and an incidental parameters problem, running non-linear models with such a large number of fixed effects is not feasible. Hence, we run these as linear probability models. In order to compare to a baseline, column 1 reports the coefficients from the linear probability model with the same set of controls as above.

We run these regressions to show how the effect of EPP varies under different identification assumptions. Take the regression with seller fixed effects. Here the identification comes from changes in an individual seller's EPP rating over time – the fixed effect controls for cross sectional variation in sellers. This regression alleviates concerns of the type that cross-sectional

Table A-4: New vs. Used

	New b/se	Used b/se
Seller Feedback Score	-2.17e-08	-0.00000107***
	1.29e-08	9.15e-08
Percent Positive Dummy excluded: $0 < .994$		
$\geq .994 < 1$	-0.0611***	-0.0280***
	0.00172	0.00485
$= 1$	-0.104***	-0.0366***
	0.00289	0.00492
EPP	1.134***	0.819***
	0.00815	0.0153
Item Price	-0.00196***	-0.000777***
	0.0000164	0.0000292
Seller Standards Dummy excluded: Below Standard		
Standard	-0.0757***	-0.119***
	0.00468	0.00969
Above Standard	-0.0677***	-0.0998***
	0.00394	0.00956
ETRS	-0.0843***	-0.137***
	0.00397	0.00954
Seller Number of Trans	1.62e-08*	0.000000554***
	7.56e-09	4.71e-08
Constant	-0.432***	-0.254***
	0.00762	0.0154
N	9,669,511	1,951,384

Regression includes controls for (coefficients not displayed) auction type (auction, fixed price), item category, and the transaction number of the buyer at the time of the focal observation.

Table A-5: Fixed Effects Regressions

	OLS b/se	Seller Fixed Effects b/se	Buyer Fixed Effects b/se
Seller Feedback Score	7.30e-10	-0.000000122***	-7.34e-09***
	1.07e-09	5.09e-09	9.76e-10
Percent Positive Dummy excluded: $0 < .994$			
$\geq .994 < 1$	-0.00875***	0.000774*	-0.000930***
	0.000187	0.000377	0.000163
$= 1$	-0.0118***	0.00683***	-0.00200***
	0.000281	0.000637	0.000242
EPP	0.130***	0.421***	0.0287***
	0.000804	0.00285	0.000749
Item Price	-0.000302***	-0.000272***	-0.000136***
	0.00000181	0.00000284	0.00000178
Seller Standards Dummy excluded: Below Standard			
Standard	-0.00744***	0.00402***	-0.00129**
	0.000471	0.000686	0.000404
Above Standard	-0.00751***	-0.00299***	0.000240
	0.000415	0.000505	0.000355
ETRS	-0.0112***	-0.00308***	-0.00104**
	0.000418	0.000529	0.000361
Seller Number of Trans	-3.18e-09***	1.08e-08***	3.73e-09***
	5.71e-10	1.95e-09	5.28e-10
Used / New Dummy excluded: New			
Refurbished	-0.00228***	0.00260**	0.00122*
	0.000530	0.000804	0.000475
Used	-0.00113***	0.00465***	0.000947***
	0.000228	0.000452	0.000216
Constant	0.448***	0.269***	1.064***
	0.000788	0.00193	0.000767
N	11,883,455	11,883,455	11,883,455

Regression includes controls for (coefficients not displayed) auction type (auction, fixed price), item category, and the transaction number of the buyer at the time of the focal observation.

variation in product quality (or product type) correlates with EPP, and this correlation generates buyer exit, not the direct effect of seller quality on buyer exit. Given the positive coefficient on EPP, one would have to believe that sellers change their mix of products in concert with changes in their EPP score for this regression to be confounded, a story we find implausible.

The regression with buyer fixed effects controls for an individual's native propensity to exit regardless of transaction quality and is identified based on changes in a buyer's propensity to transact with sellers of different EPP levels over time rather than cross-sectional variation in different buyers' propensity to transact with different sellers. Thus, for this regression to be confounded, one would have to believe that buyers shift their selection of sellers in concert with their probability of exit for reasons that are unrelated to transaction quality.

We note that while we believe these regressions provide substantial evidence for the exogeneity of EPP, further unassailable evidence comes from the randomized field experiment documented in Section 6 of the paper.