THE VALUE OF REPUTATION ON EBAY: A CONTROLLED EXPERIMENT*

Paul Resnick

Richard Zeckhauser

John Swanson

Kate Lockwood[1]

The latest version of this working paper can be found at http://www.si.umich.edu/~presnick/papers/postcards/

This is the final pre-publication version, January 3, 2006. The paper is forthcoming in Experimental Economics.

Abstract

We conducted the first randomized controlled field experiment of an Internet reputation mechanism. A high-reputation, established eBay dealer sold matched pairs of lots -- batches of vintage postcards -- under his regular identity and under new seller identities (also operated by him). As predicted, the established identity fared better. The difference in buyers' willingness-to-pay was 8.1% of the selling price. A subsidiary experiment followed the same format, but compared sales by relatively new sellers with and without negative feedback. Surprisingly, one or two negative feedbacks for our new sellers did not affect buyers' willingness-to-pay.

I Introduction

As the Internet grows as a means of executing transactions, each buyer's array of possible vendors is mushrooming. On auction sites, like eBay, users already buy and sell things from others across the nation and around the world. Despite opening many new venues, this electronic bazaar puts stress on some of the foundations of the traditional market place. In traditional markets, a buyer can usually "squeeze the orange", e.g., inspect the vintage plate, before buying. Beyond this, a seller's reputation, possibly built over many years, including the cost of her physical facility and her standing in the community, provides a useful signal about the seller’s ability and keeps her honest and diligent.

Sales over the Internet lack these tools of reputation. No seller has long been in the electronic market. A seller's physical facility may be her kitchen, and virtually no buyer knows a seller's standing in the community. To be sure, some sellers, such as Dell and L.L. Bean, borrow reputations from elsewhere. However, for tens of thousands of sellers, there is no outside instrument of reputation. In such circumstances, the temptation to sellers to misrepresent products, e.g., exaggerate their quality or misrepresent their provenance, are great. So too is the temptation to sloth, to ship slowly or sloppily after receiving payment.[2] This should lower the price that buyers are willing to pay, since they are forced to assume some risk for the quality and utility of the good being traded. Unless sellers can provide sufficient information about product quality and their own quality in transaction fulfillment, low-quality products and sellers will drive out those of high quality and the market will shrivel (Akerlof 1970).

Internet marketplaces have been forced to find a substitute for traditional seller reputations. Important systems have been introduced to enable the systematic elicitation and distribution of reputational information. These systems collect information on the past behavior of a seller, or for that matter of a buyer, and then make that information available to potential future transaction partners. Because people know that their behavior today will affect their ability to transact in the future, not only with their current partner but with unknown others as well, opportunistic behavior is deterred. Moreover, less reliable players are discouraged from joining the marketplace. Reputation systems seek to inform buyers about whether potential trading partners are trustworthy, and thereby to make chiseling and cheating rare and losing propositions.

Though disadvantaged in the respects described above, Internet markets also have significant advantages in establishing reputations. First, any information that is gleaned can be near costlessly tallied on a continuing basis, and written assessments can readily be assembled. Second, that information can be near costlessly transmitted to millions of potential customers. (By contrast, word of mouth distribution loses vast amounts of information, with different buyers hearing significantly different assessments of the same seller. It also entails a per-telling cost.) Third, the Internet has the potential, though at present not the reality, for sophisticated processing of information, e.g., using Bayesian calculations, and for using micropayments to induce careful and honest assessments from transactors (Avery, Resnick and Zeckhauser 1999; Miller, Resnick and Zeckhauser 2005).

The same factors that advantage the Internet in establishing reputations make it a wonderful place to study the role of reputations. In auction markets, such as eBay, where the vast majority of sellers are otherwise unknown, the researcher can get precise measures of current reputation. Moreover, that information is cheaply available to all participants in the marketplace. The contrast with traditional markets is stark. A retailer in a community, say, may have a strong reputation with some individuals, a weak one with others, and his reputation may be unknown to newcomers. Hence, any reputation measure would have considerable noise. With Internet auction sites, reputations are common knowledge, or at least commonly available knowledge. This paper assesses the value of reputations, capitalizing on the known-reputation feature of the Internet auctions conducted on eBay, and the ready availability of data on the frequency and price of sales there.

There have been a number of observational studies of eBay reputations and sales. (Fifteen are addressed in section II.B.) Observational studies are limited in two ways. First, they can only examine the impact of reputation in markets for standardized goods, since they depend on naturally occurring variation in reputation for sales either of identical goods or of goods whose outside market value can be used as a control. Second, such studies are invariably plagued by an omitted variables problem. In the eBay context, for example, we can't tell whether a seller is getting a higher price because of a better reputation, or alternatively, a more attractive web site, superior presentation of items, better answers to inquiries, or superior merchandise. Except in rare instances, observations on these alternative explanations are not immediately detectable by the researcher. Moreover, both reputation and omitted measures of a seller's skill or merchandise are likely to be correlated with a seller's experience level. Thus, effects of the omitted variables are likely to be attributed to the reputation, leading to an overestimate of the true impact of reputation.

Laboratory experiments can more clearly isolate the effects of reputation. Keser (2003) utilized an “investment game” where one player’s trust increases the total payoffs but leaves her vulnerable to the other player taking an unfair portion. When subjects who had not previously interacted with each other were informed of each other’s past play, both trust (investment) and trustworthiness (return of profits to the trustor) were higher. Bolton, Katok, and Ockenfels (2003) utilized an analogous two-stage game where buyers decide whether to send money, and sellers then decide whether to ship the item. In the reputation condition where subjects were informed of each other’s past play, trust and trustworthiness increased as compared to a no reputation condition, but still did not reach the levels found in a repeated-play condition.

In recent years, field experiments have gained favor, as a way of combining the controls of lab experiments with the external validity of studying behavior in natural settings. For example, Camerer manipulated betting markets (Camerer 1998), List and Lucking-Reiley compared different methods of soliciting charitable contributions (List and Lucking-Reiley 2002) and of auctioning multiple units of a good (List and Lucking-Reiley 2000), and Lucking-Reiley tested the effects of auction formats (Lucking-Reiley 1999), minimum bids (Reiley forthcoming), and hidden reserve prices (Katkar and Lucking-Reiley 2000).[3]

We conducted the first randomized controlled study of the value of eBay reputations in the natural setting of actual eBay auctions. We were fortunate to secure the collaboration of a highly experienced eBay postcards seller (Swanson). He prepared matched pairs of auction lots, and a random device then determined which lot was sold under his established, extremely high reputation identity, and which lot was sold through newly created, unknown sellers. Roughly similar web sites were prepared for all sellers. Other factors, such as shipping, billing and payment procedures were constant across our sellers.

Section II describes the eBay reputation system and examines prior research in more detail. Section III describes our randomized controlled experiment and its results. Section IV discusses the results and discusses their implication. Section V concludes.

II The eBay Reputation System

There are many sites with reputation systems of some sort. The eBay system is undoubtedly the biggest and best known. eBay has millions of items available for bid at any time. The eBay reputation system enables users to leave feedback about interactions with each other. The system is transaction based: to leave feedback for each other, two users must actually have completed an auction. After the auction ends, the buyer and seller each have the opportunity to rate each other's performance with either a 1 (positive), a 0 (neutral), or a –1 (negative). Users also have the opportunity to leave a one-line text comment, and rated individuals can respond to comments that they feel were unfair. Users’ net reputation scores are calculated as the count of distinct users who gave positive feedback minus the count of those who gave negative feedback. The seller’s net reputation score -- positives less negatives -- is automatically displayed on the auction page for each item she lists. Hence, potential buyers see this rating before bidding.[4] A buyer can choose to click on the net score in order to see a more detailed breakdown into positive, negative, and neutral over a series of time periods. The buyer can then scroll down to see individual comments, with the most recent ones shown first. A user who is new to the system starts with a net feedback score of zero and has a sunglasses icon displayed next to his or her screen name for the first 30 days of membership. Users may change their eBay identities by registering again, but must then start all over as new users with a zero reputation score.

II.A Expectations About and Experience with Feedback and Reputations on eBay

We focus on seller reputations, since those for buyers matter little. (The seller can simply wait to get paid and thus incurs little risk.) If buyers are uncertain about seller trustworthiness, they will reward better seller reputations by raising their offers, even though each buyer is only concerned about his own welfare. Indeed, if it is costly to maintain a reputation for high quality, then a good reputation needs to be rewarded by at least the cost of building one. A bad reputation or a decline in reputation should incur a loss that exceeds the benefit from opportunistic behavior (Shapiro 1983). Thus, in equilibrium, a good reputation must command a price premium.[5] Since sellers who get negative feedback can start over relatively easily, buyers need to impose some disadvantage on sellers with no feedback at all (Friedman and Resnick 2001).[6] Finally, we should expect that buyers will not provide information to help determine seller reputations, since to do so incurs a cost, and free riding is hard to punish.[7]

eBay reputations in practice do not illustrate pure rational game-theoretic processes in action. (Resnick and Zeckhauser 2002), hereafter RZ, found that even though the incentive to free-ride is clear, half of the buyers on eBay provided feedback. This suggests that a high level of courtesy exists on eBay. After a satisfactory transaction, you provide a relatively low cost positive feedback just the way you provide a thank you in everyday discourse.

The most striking feature about eBay feedback is that it is so positive. Sellers received negative feedback only 1% of the time, and buyers 2% (RZ). Given their rarity, negatives should be much more consequential than positives in affecting a seller's overall reputation. The specifics of negatives should be much more informative. However, eBay offers no search mechanism to find negatives.

For a seller, what constitutes a good reputation in the eBay feedback system? The answer depends on how buyers behave. There are many possibilities. At one extreme, buyers may in effect be Bayesians, effectively incorporating information not only from reputation scores but from a seller's product, geographic location, written comments, whether she has an expensive website, etc. Such buyers would be to statistics as Molière's Monsieur Jourdain was to prose, unknowing but effective users. At the opposite extreme, buyers may employ simple heuristics that are far from optimal, as human decision makers are known to do (Tversky and Kahneman 1974).

II.B Prior Empirical Studies

In the last few years, a large number of empirical studies of the effects of eBay’s reputation system on sales have been undertaken. We are aware of at least 15, as summarized in Table 1.[8] One study, BP, follows the logic of a lab experiment, splicing different seller reputation scores into real auction listings and asking subjects to indicate how much extra they would pay to different sellers. The remaining studies all follow the logic of hedonic regression, though their details vary in important ways. Each is an observational study of a set of items whose sellers had varying reputations. Each correlates the reputations with auction outcomes, while controlling for possible confounds. Mean prices for the items studied ranged from $33 to $1621. Generally, these are relatively high-priced items for eBay: the median selling price for an item in a dataset analyzed in RZ was less than $15.

The results are broadly consistent with the theoretical expectation of buyers paying more to sellers who have better reputations, but the results do not yield a consistent picture. While most studies find some effect of positive feedback (or net feedback), three do not (LBPR, E, CH). At the larger end of effect sizes for positive evaluations, the model in L finds that sellers with more than 675 positive comments earned a premium of $45.76, more than 10% of the mean selling price, as compared to new sellers with no feedback. Similarly, while most studies that examined negative feedback found an impact on probability of sale or price, L did not, E found an effect only for established sellers, CH found an effect only after eBay changed its display in 2003 to show the percentage negative along with the composite score, and MS found that more negatives actually increased the number of bidders. At the larger end of effect sizes for negatives, LBPD, looking at collectible coins, finds that a move from 2 to 3 negatives cuts the price by 11%, about $19 from a mean price of $173.

These studies all control in some way for variability in underlying product values.[9] Without such controls, as LBPD demonstrate, omitted variable bias can produce misleading results, including positive reputation scores that appear to lower prices while negative reputations raise them. Two methods have been used to control for value of the product, so as to correct this omitted variable bias. One is to study auctions of identical products, such as computer processors (HW), specific collectible items (MS, RZ), or coins whose value comes from their gold content rather than their collectible value (MA). An alternative approach is to include book value or market price in regressions to control for the differences in item values (LBPR, BH, DH, BP, KM, JK).

Many auctions do not result in transactions, and frequently items do not even receive bids. As expected, reputation affects the probability of sale (E, JK, MA, L) as well as price.[10] It also affects the probability of individual bidders entering the auction (BH), and the number of bids (BH, MS). It also affects buyers' subjective assessments of a seller's trustworthiness (BP).

Many studies include the number of bids as a control variable in regression models. However, LBPD argues, and we agree, that the number of bids is an endogenous indicator of the impact of reputation on price, and should not be an independent variable in a simple regression model. Number of bids or bidders can be treated as an outcome variable (BH), or may be modeled simultaneously as both an effect of reputation and an independent contributor to price (MS).

The observational studies surveyed here rely on being natural quasi-experiments. As such, the danger remains that unknown or otherwise unmeasured covariates of reputation produce outcome differences that will be mistakenly attributed to the reputation score. Most studies control for covariates that are easily coded from auction listing, including whether a picture of the item is displayed, whether credit cards are accepted, time of day and day of the week that the auction closes, and length of the auction.[11]

Despite these attempts to control for potential confounds, however, some important ones remain. First, private email communications between buyer and seller during the bidding can influence the buyer's willingness to bid high, but such communication is invisible to the researcher who can monitor only the eBay website. Thus, for example, sellers who are very responsive to email inquiries may have induced buyers to bid higher, and this responsiveness may be correlated with reputation. Second, the completeness and quality of the seller's description page, and the aesthetics of its layout may affect bidding. Since many buyers type in search terms to find auction listings, including the right words in an item's title, and spelling it correctly, can make a huge difference (Schemo 2004). More experienced sellers, who have more feedback, are also more likely to have learned how best to describe their items. In principle, a researcher could code for the quality of listings, but in practice it is very difficult to do so. A third and related confound is that more experienced sellers may tend to have higher quality goods, and the difference may be apparent to buyers but not easily coded by researchers.

A controlled field experiment, as presented in this paper, offers two main advantages over observational studies, yet maintains the external validity of observing the behavior of buyers in a natural market setting.[12] First, it automatically controls for confounds. We varied the seller's reputation without varying his responsiveness to emails from potential bidders, his skill at listing items, or the quality of his goods. Second, it makes it possible to investigate the impact of reputation in markets for non-standardized goods for which book values are unavailable. A controlled experiment requires only one matched item against which to compare an auction outcome rather than requiring a comparison against the market price data for a large set of similar items.

III The Randomized Controlled Experiment

III.A Procedures

We worked with an established eBay dealer with a high reputation (net score above 2000 as of the beginning of the study, with just one negative). In real life, that reputation belongs to coauthor John Swanson, who runs a business with Nina Swanson dealing primarily in vintage postcards. They typically list dozens of items for sale each week on eBay. (They also sell items live at postcard shows and other events.)

In addition to selling items using his established identity with a strong reptutation—hereafter often referred to as STRONG--, our dealer created seven new eBay seller identities. Each new seller started with no feedback. Each of the new sellers will be referred to generically as NEW. The dealer provided the same great service (communication, packaging, shipping) when listing items under any of the seller identities, but buyers looking only at the item listings and seller information seen on eBay would not have known that the same high-quality dealer was behind all the sellers.

Our primary experiment compared listings of 200 items by STRONG to listings of matched items by one of the new sellers. Information on results was gathered directly from the eBay webpage, using a spider (automated computer program) to collect data. At the completion of each auction, detailed information was collected about the bids placed, the selling price of the items, and the feedback of both the buyer and the seller at the time of the auction. This data was then double checked against the records kept by our seller.

Since the postcards were not standard items, the matching, which was done on both subject and value, required the dealer's judgment. To avoid possible bias, a random device determined which item in each pair would be listed by STRONG and which by one of the new sellers. By matching items rather than trying to control for variation in item value, the controlled experiment allows us to examine the effects of reputation on sales of unique, used items, in our case vintage postcards. For such items, there is a great deal of information asymmetry between seller and buyer about item condition, and no established book values to guide buyers.

The typical item was a “lot” of vintage postcards, titled something like “Vintage Valentine Postcards (36)”, where the number in parentheses indicated the number of cards in the lot. The item description indicated the general condition of the cards and provided photos of one or a few cards in the lot. The dealer followed his usual practices to determine a minimum starting bid for each item. The range was $4.99-$49.99 with a mean starting bid of $13.13 and a median of $9.99. Informal analysis suggests that listings of this value are quite typical on eBay, not only in the vintage-postcard category, but overall.

Over the course of the experiment, which took twelve weeks, five of the new sellers presented 20 lots each for sale, and two new sellers presented 50 lots each. This was done to allow for accumulation of different amounts of feedback. The lots, which could include multiple items, were divided into five sets, grouped by listing price. Each set contained 20 pairs (40 lots total) and was listed in two different weeks, which in turn were separated by a week in the middle to ensure that there would be no overlap in the availability of the matched items. For half of the pairs in any set, again determined at random, STRONG listed its lot in the first week and for the other half, the new seller listed its lot first. Two weeks later, the other seller listed the other lot of the pair. In any set, 10 lots each were listed by the two high volume new sellers, and 4 by each of the five low volume new sellers.

Table 2 illustrates how this counterbalancing worked. For example, the first set of 20 pairs had starting (minimum) prices in the $9.99-14.99 range. Half of each pair was listed in week one and the other half in week 3 of the experiment. For half of the pairs (the 1B group), the new seller listed its lot in week 1 and STRONG listed its matching lot in week 3. For the other half, the order was reversed. The starting bids for the lots were balanced among the seven new sellers as closely as possible.

Several steps were taken to make it difficult for buyers to notice that an experiment was underway, and we received no communications suggesting that any bidders noticed lot pairings or other elements of an experiment. Lots were listed in a category that typically has thousands of lots for sale. Each of the new sellers used a slightly different format for listing lots (e.g., “36 Valentine Postcards, Vintage”). Each of the new sellers had its own e-mail address from which correspondence emanated. Finally, the two halves of each matched pair were listed for sale in different weeks rather than at the same time. Care was taken to assure that each seller listed the lot using the same information (selling price, tax, shipping cost, payment methods and description), while maintaining a unique look and feel to its listings, as illustrated in Figures 1 and 2. The listings were created using AuctionHelper, the program our dealer uses to list his lots under his established identity. By giving each seller a unique look, we were able to avoid making it apparent that all sellers were being operated by the same person. However, by including all the same information, we kept the listings matched as closely as possible.[13]

Each of the new sellers began with no prior feedback. During the course of the experiment, feedback from buyers in previous time periods became visible. Our high volume sellers ended up with 12 and 17 positive feedbacks and our low volume sellers ended up with between 5 and 14 positive feedback points, with an average of 9.2 positive points. None of our sellers received any negative feedback from their sales.

In a second experiment, we tested the effects of negative feedback. Such feedback, being scarce, carries much more information content. Especially in a brief reputation, it should hurt sellers if noticed. To prepare for the second experiment, we purchased lots from three of the new sellers in order to give them each one or two negative comments. The experiment, lasting three weeks, compared results for new sellers with and without negative feedback.

A previous study analyzed reasons for negative feedbacks at eBay, and found explanations ranging from minor (slow shipping) to major (sellers who cashed checks but never sent lots) (Resnick and Zeckhauser 2002). Matching one of the common complaints found in the earlier study, all of the negative feedback comments we left indicated either that the item received was not as described or was in worse condition than in the auction listing text, though we purposefully did not provide any details. In all cases the user that gave the feedback was itself a fictitious entity with zero feedback. The "item not as described" negative highlights one of the big problems with Internet auctions: the inability of potential buyers to decide whether to trust the seller's description of an item. Figure 3 shows the feedback profile for one of our sellers at the beginning of the second experiment. In all cases, negative feedback was displayed at the top of the comments at the beginning of the second experiment. Feedback accruing during the second experiment pushed the negative comment(s) down, but never off the first page.

With our negative feedbacks given, new sellers with similar amounts of positive feedback were paired, as shown in Table 3. The dealer then listed 35 more matched pairs of lots, with half of each pair listed by a seller with no negative feedback and the other listed by a seller with negative feedback. As in the primary experiment, lots were listed in separated weeks, with order of listing counterbalanced between the two sellers. Randomization was used as before. The two high-volume sellers from the primary experiment were paired together. The next pair was formed from the two new sellers with the next most feedback, and the last pair was the two sellers with the next most feedback after that. The new seller that ended the primary experiment with the lowest feedback score (4) was not used in the second experiment.

III.B Hypotheses

The established seller has a vastly better reputation than any of the new sellers. The main hypothesis of our study is that buyers will view the established seller as less risky and thus will pay more.

H1: Buyers are willing to pay more to a seller with a strong positive reputation (STRONG) than sellers without such a reputation (NEW).

We expected that buyers would give little weight to the one negative feedback of the established seller, given the positive feedback from more than 2000 distinct buyers. However, in the second experiment, when sellers with relatively few positive feedbacks also have negative feedback, we expect buyers to distrust them.[14] Our subsidiary hypothesis is parallel to H1 for the primary experiment.

H2: The new sellers with negative feedback will reap lower profits than those without negative feedback. The new seller with two negative feedbacks will reap lower profits than those with just one negative feedback.

III.C Results

To analyze the impact of reputation on buyers' willingness-to-pay, we compare outcomes for matched lots, one listed by STRONG and one by a new seller. Our statistical tests are conducted putting the results for all the new sellers together.

Buyers' collective willingness-to-pay for a lot through an ascending auction is imperfectly observed in eBay auctions.[15] The ideal situation for the experimenter is when both items in a pair sell, at higher than their minimum bids. When neither STRONG nor NEW sells a lot, i.e., the opening (minimum) bid is above any buyer's willingness-to-pay, we get little information about buyers' relative willingness-to-pay for the lot from the two sellers. When one sells but not the other, the observation provides a lower or upper bound on the ratio of buyers' willingness-to-pay. For example, if only STRONG sells a lot, then buyers were willing to pay somewhat less than the minimum bid to NEW, and hence the ratio of willingness-to-pay by the top bidder was at least the price paid to STRONG divided by the minimum bid.

Given this limited information, our tests for significant differences begin with nonparametric tests. We then examine whether there are significant differences in the number of non-sales. Finally, we employ censored regression techniques to produce parametric estimates of the magnitudes of differences, taking into account the censored observations that result from non-sales.

We expect the ratio of buyers' willingness-to-pay to be lognormally distributed. That is, $\ln\frac{STRONG\_price}{NEW\_price} = \ln(STRONG\_price) - \ln(NEW\_price)$ should be normally distributed. Our hypothesis is that the mean is greater than 0.

Test of H1: Table 4 shows the results of a sign test on the paired differences between $\ln(STRONG\_price)$ and $\ln(NEW\_price)$. When neither lot sold, we exclude the lot, as no information is available about buyers' relative willingness-to-pay. When STRONG sold a lot but NEW did not sell the matched lot, the sign of the difference was positive; when the reverse occurred, it was negative. Of course, if both sellers sold a lot, then the observed difference was used. The setup of the auction, with identical starting prices and increments for a pair, tended to produce ties (e.g., if both sales had just one bidder, both sales were exactly at the minimum bid), which worked against any hypothesis of difference. A one-sided non-parametric sign test was significant: Pr(#positive >= 81 | median of the differences = 0) = Binomial(n = 139, x >= 81, p = 0.5) = 0.0308.

There are two ways STRONG can outperform NEW: sell more lots at or above the starting price, and secure a higher price for lots sold. Table 5 shows the results on frequency of sale. A chi-square test concludes that the probability of sale was not independent between the two sellers (p<.001). The NEW sellers were much more likely to sell when STRONG sold and vice versa; this correlation suggests that our dealer did a good job of pairing lots. That the correlation is far below 1 indicates that other factors, including chance, often determine whether one, both or neither of a matched pair of postcard lots gets sold. That some lots sold and some did not indicates that the dealer could not perfectly guess the appropriate starting bid.[16]

Overall, the strong seller listings sold 63% of the time, the new seller listings 56% of the time. A one-sided sign test on the difference between STRONG_sold and NEW_sold approaches significance, Pr(#positive >= 40 | mean of differences=0) = Binomial(n = 67, x >= 40, p = 1/2) = 0.0710.

Next, we employ parametric techniques to estimate the magnitude of the difference between buyers' willingness-to-pay for lots listed by STRONG or NEW. One conceivable approach would be to consider only lots that sold both times. This would have two disadvantages. First, it would reduce the sample size quite a bit, reducing the statistical power. Second and more important, it would introduce a truncation bias, since the new sellers sold fewer lots. Thus, the observations of sold lots for NEW reflect more extreme points than for STRONG in their respective distributions of buyers' willingness-to-pay.

To avoid this, we employed a censored normal estimation that makes use of the pairs when either or both of the sellers sold the lot (151 of the 198 cases). Employing this method, the estimated mean difference between ln(STRONG) and ln(NEW) is .078, as shown in column one of Table 6. This is positive in accord with H1 and the one-sided test is significant (p=.044).[17]

We conducted additional censored normal regressions to test for possible confounds. One potential confound is that the difference may depend on the value of the lot being sold. For example, buyers may be more willing to accept risks on lower priced lots. As shown in Column 2, however, there was no significant difference due to the minimum bid. Column 3 checks whether the counterbalancing was effective: there was no significant difference between lot pairs where STRONG sold first or second. Column 4 includes fixed effects of the new sellers (perhaps, despite our efforts to homogenize, some new sellers had more attractive names or formats for listing lots), with one of the new sellers the omitted variable. None of the individual coefficients is significant and the chi-squared test on the likelihood ratio indicates that the different-attractiveness model as a whole is not significant.

Since the large amount of feedback of the established seller made a difference, it is natural to investigate whether the few positive feedbacks that the new sellers accumulated during the primary experiment were sufficient to alter buyers' willingness-to-pay. The evidence says no. Column 5 of Table 6 compares sales in the last four weeks to those in the first four weeks. NEW actually fared slightly better relative to STRONG in the first four weeks than in the last four, though the difference did not approach significance. Column 6 tests directly whether the amount of prior positive feedback for the NEW seller mattered. Again, the coefficient has a sign opposite to that expected, but again it is not significant.

We now turn to analysis of the secondary experiment. Surprisingly, **H2**, which posits that negatives in a brief reputation will hurt revenues, was not confirmed. The sellers without negatives sold 16 of the 35 lots (46%) and those with negatives sold 14 lots (40%). The difference is far from significant given the small sample size. Moreover, the lots listed by the sellers without negatives more often received lower prices when they did sell: the sign of the difference in willingness-to-pay favored the sellers without negatives 9 times, the sellers with negatives 11 times. A censored normal estimation also shows no significant difference between buyer willingness-to-pay to the two types of sellers.

IV Discussion and implications

Our basic finding is that, in conformance with the theory of reputations and with lore about eBay, the market rewards (pays more to) a seller who has accumulated a lot of positive feedback. The estimated mean difference between ln(STRONG) and ln(NEW) is .078. Exponentiating .078 yields 1.081. Thus, we find that buyers are willing to pay 8.1% more for lots sold by STRONG than NEW.[18] Given that the logarithm is a concave function, the estimated difference at the mean of the distribution may not exactly reflect the mean price ratio calculated using straight dollar prices. However, taking logarithms does not affect the median. The actual median of the price ratios occurred for a pair that sold under both NEW and STRONG. STRONG received 7.2% more for this lot, which is close to our 8.1%.

Similar findings from previous observational studies could conceivably have attributed to reputation scores effects that were really due to unmeasured differences in listing quality, product quality, or seller responsiveness to inquiries of potential bidders. Our experiment controlled for these potential confounds. Before concluding with certainty that the reputation score itself was responsible for the market's greater willingness to pay, however, it is worth considering some threats to validity that remained even in our controlled experiment.

First, there may have been differences in the quality of listings, even though the same seller created them. To avoid tipping our hand that an experiment was in progress, each seller used a slightly different look and feel. Since there were not statistically significant differences among the new sellers when fixed effects were included in the censored normal regression, we believe that the dealer was able to present the same information in multiple ways without having a large impact on the market. Extrapolating, it seems unlikely that the difference between the established seller and the new sellers was due to the presentation of listings, but we cannot tell for sure.

A second threat to validity could come from repeat customers. If some previous customers of the seller with the strong reputation were more likely to look at that seller's listings, or to base their assessments of a lot on their personal experience with that seller, the observed difference in outcomes might be attributable to previous customers rather than to the public reputation. We do not have a list of all previous customers of STRONG, but we were able to harvest a list of all user ids of previous buyers who left feedback. This is sufficient to estimate whether bidders for STRONG were more likely than bidders for NEW to include previous buyers from STRONG. In fact, 23 of STRONG's 137 distinct bidders had given feedback to STRONG before the experiment started, accounting for 24% (44 out of 181) of the total distinct lot-bidder combinations. On the other hand, 13 of the NEW sellers' 121 distinct bidders had given feedback to STRONG, and they accounted for only 11% (15 out of 136) of the total distinct lot-bidder combinations. Thus, we cannot rule out the possibility that the private reputation with specific customers rather than the public reputation embodied in the sellers' feedback profiles was responsible for some or all of the observed difference between STRONG and NEW.

A third and related threat to validity could come from multiple purchases. A buyer, having once bid on a lot from a seller, may look for other listings by that seller.[19] STRONG had more lots listed at any one time than any of the NEW sellers, for two reasons. First, there were multiple NEW sellers, but all were paired with the same

STRONG seller. Second, our dealer conducted additional sales during the time of the experiment using his regular selling identity, the STRONG seller in the experiment. If there were no such search effect, then bidders on STRONG and NEW lots should have been equally likely to bid soon after on a lot listed by STRONG. To test this null hypothesis, we examined our relatively complete bidding records[20], restricting our attention only to the bidders who had not purchased from our dealer prior to the experiment. For each bidder's first bid on a lot, on both STRONG and NEW listings in the experiment, we calculated whether the bidder also bid within 3 days on any other STRONG listing (whether part of the experiment or not). For bidders on STRONG lots in the experiment, 13 out of 137 (9%) also bid on some other STRONG listing. For bidders on NEW lots, 10 out of 121 (8%) also bid on some other STRONG listing.[21] Thus, there is little evidence to suggest that the better market performance for STRONG derived from a larger volume of listings. Still, an ideal future field experiment should have different sellers list the same number of lots. Future observational studies of the eBay market should also add as a control variable the number of simultaneous listings for the seller.

The results from our second experiment seem somewhat puzzling. We found no impact of negative feedback. This could merely be a spurious inability to detect a small effect with a small sample size, though the trends did not even point toward an effect.

Here the threats to validity described above did not come into play, since all sellers were new, had closely matched previous selling histories, and listed the same number of lots. If the effects in the primary experiment were due mainly to private reputations and volume of listings, the lack of a measurable difference in the second experiment would not be surprising.

Alternatively it is possible that the public feedback profiles matter, but that buyers discounted the negative feedback because it was terse and did not explain exactly what was wrong with the lot, or because the feedbacks were given by buyers who themselves had no feedback. For the seller with two negatives, buyers may have suspected something fishy because the two negatives, both from buyers with no feedback themselves, were both posted within a minute of each other (in retrospect, an experimenter's error).

Another possible explanation along these lines is that the market treats new sellers as untrustworthy but does not distinguish among feedback profiles of new sellers. Perhaps buyers are not aware of the predictive value of an early negative on future problematic transactions, think one or two negative feedbacks may be an aberration, and thus pay attention to the percentage of negative feedback only for sellers with longer histories. One piece of evidence against this hypothesis is that Dewally and Ederington (2006) found no evidence that buyers give less credence to the percentage negative when it is based on just a few feedbacks, and that buyers place much more weight on the first few feedbacks for a seller than they do on their priors about sellers in general.

Our primary conjecture as to why negative feedback had no measurable effect is that most buyers simply did not bother to click through to look at the detailed feedback, and instead merely relied on the overall score (number of positives minus number of negatives) that eBay displayed as part of the lot listing. This value barely differed among the seller pairs. In support of this conjecture, Cabral and Hortascu (2004) found that the percentage of negative feedback did not affect market price until after eBay's 2003 transition to displaying the percentage of negative feedbacks in addition to the composite score.

While the field experiment methodology provides a more controlled test of whether reputation score affected the seller's revenue than can an observational study, it leaves many questions unanswered. In addition to our lack of control for private reputation and the search effects of listing volume, it is not clear how well these results would generalize to other categories of lots on eBay, or to other marketplaces that employ reputation systems.

Even for the vintage postcard market, our experiment provides insufficient information to determine whether the market is in an equilibrium that can be explained with a reputation model based on some combination of moral hazard and adverse selection, such as that in Diamond (1989) and Cabral and Hortascu (2004). The first condition for such equilibria would be that the price premium that buyers pay to sellers with better reputations should be calibrated to the reduced risks they face. There are several difficulties in determining whether that condition is met in eBay's vintage postcard market. First, since we have estimated only the market's willingness to pay, and not a structural model of individual willingness to pay, we do not know how individual buyers are discounting based on the risks they perceive. Second, although we have estimates of how frequently problems occur, we do not know how much buyers gain from successful transactions or lose from problematic transactions.

The other conditions for reputation equilibria involve sellers. To deter moral hazard, the cost of building and maintaining a high-quality reputation (including the lost profits from not cheating) should be just enough to balance out the profit premium such a reputation secures. If quality is a long-term property of the seller and not a matter of strategic control on each sale, then it must be profitable for all the desirable types, and only them, to participate in the marketplace. To assess whether the market is over- or under-rewarding reputations, however, would require estimates of the profit that could be made on each transaction from cheating, the number of transactions required to build up a strong reputation, and the number of times a seller could cheat before destroying a strong reputation. None of these is available from the experiments reported here.

V Conclusion

The eBay Feedback Forum illustrates Yhprum's Law.[1] Systems that shouldn't work sometimes do, or at least work fairly well. At least prior to eBay's recent change toward displaying the percentage of negative feedback more prominently, buyers appeared not to pay sufficient attention to negative feedback for relatively new sellers. That should have created an opportunity for sellers to cheat and profit. More generally, it is not clear whether the price premiums reflect a reputation equilibrium.

Nonetheless, we find that eBay's public reputation scores play a significant role in the marketplace, and that virtually all significant sellers have strong reputations. In our controlled experiment, a seller with a strong reputation received a price premium, even holding constant quality of goods, skill at listing, and responsiveness to inquiries, all potential confounds in previous observational studies. Looking at matched pairs of lots – batches of vintage postcards – buyers were willing to pay a STRONG-reputation seller

[1] Yhprum is Murphy spelled backward. Murphy's Law is a popular adage stating that anything that can go wrong will go wrong.

8.1% more on average than a NEW seller. While the experiment did not control for the potential confounds of private reputation information and volume of seller listings, it did rule out several potential confounds for previous observational studies, including seller skill, unmeasured but observable product quality, and seller responsiveness to bidder inquiries.

Perhaps the Feedback Forum works because the entry costs in time and skill to become an eBay seller discourage low-integrity sellers from joining the market. Perhaps the skilled yet unscrupulous can make more money elsewhere. Perhaps the price premium paid to sellers with strong reputations, while small on any individual transaction, is sufficient to maintain their good behavior both because it can be enjoyed many times once a reputation has been built and because it can be destroyed fairly quickly. Perhaps a fairly effective reputation system is good enough: A reputation system that merely reduces the lure of cheating and chiseling without eliminating it may be sufficient to activate Yhprum's Law.

VI References

Akerlof, G. A. (1970). "The Market for "Lemons": Quality Uncertainty and the Market Mechanism." The Quarterly Journal of Economics **84**(3): 488-500.

Avery, C., P. Resnick and R. Zeckhauser (1999). "The Market for Evaluations." American Economic Review **89**(3): 564-584.

Ba, S. and P. A. Pavlou (2002). "Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior." MIS Quarterly **26**: 243-268.

Bajari, P. and A. Hortacsu (2003). "Winner's Curse, Reserve prices and Endogenous Entry: Empirical Insights from eBay Auctions." RAND Journal of Economics **34**(2).

Bajari, P. and A. Hortacsu (2004). "Economic Insights from Internet Auctions: A Survey." Journal of Economic Literature **42**(2): 457-86.

Bolton, G., D. Fong and P. Mosquin (2003). "Bayes Factors with an Application to Experimental Economics." Experimental Economics **6**(3).

Bolton, G., E. Katok and A. Ockenfels (2003). How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation. Available on-line at http://ockenfels.uni-koeln.de/download/papers/BKO2.pdf.

Cabral, L. and A. Hortacsu (2004). The Dynamics of Seller Reputation: Theory and Evidence from eBay, NBER. Available on-line at http://www.nber.org/papers/w10363.

Camerer, C. (1998). "Can Asset Markets Be Manipulated? A Field Experiment with Racetrack Betting." Journal of Political Economy **106**(3): 457-482.

Dewally, M. and L. H. Ederington (2006). "Reputation, Certification, Warranties, and Information as Remedies for Seller-Buyer Information Asymmetries: Lessons from the Online Comic Book Market." The Journal of Business **79**(2).

Dewan, S. and V. Hsu (2004). "Adverse Selection in Electronic Markets: Evidence from Online Stamp Auctions." Journal of Industrial Economics **52**(4): 497-516.

Diamond, D. W. (1989). "Repuation Acquisition in Debt Markets." Journal of Political Economy **97**(4): 828-862.

Eaton, D. H. (2002). Valuing Information: Evidence from Guitar Auctions on eBay. Murray, KY, Murray State University**:** 28. Available on-line at http://campus.murraystate.edu/academic/faculty/david.eaton/workpaper0201.pdf.

Friedman, E. and P. Resnick (2001). "The Social Cost of Cheap Pseudonyms." Journal of Economics and Management Strategy **10**(2): 173-199.

Harrison, G. W. and J. A. List (2004). "Field Experiments." Journal of Economic Literature **42**(4).

Henrich, B., Bowles, Camerer, Fehr, Gintis, and McElreath (2001). "Cooperation, Reciprocity and Punishment in Fifteen Small-scale Societies." American Economic Review **May 2001**: 7.

Houser, D. and J. Wooders (forthcoming). "Reputation in Auctions: Theory, and Evidence from eBay." Journal of Economics and Management Strategy.

Hossain, T., & Morgan, J. (forthcoming) ...Plus Shipping and Handling: Revenue (Non)Equivalence in Field Experiments on eBay. *Advances in Economic Analysis & Policy*, forthcoming.

Jin, G. and A. Kato (2004). Consumer Frauds and the Misinformed: Evidence From An Online Field Experiment, University of Maryland. Available on-line at http://www.glue.umd.edu/~ginger/research/ebay-Jin-Kato-Feb04.pdf.

Kalyanam, K. and S. McIntyre (2001). Return on Reputation in Online Auction Markets. Santa Clara, CA, Santa Clara University. Available on-line at http://business.scu.edu/faculty/research/working_papers/pdf/kalyanam_mcintyre_wp10.pdf.

Katkar, R. and D. Lucking-Reiley (2000). Public Versus Secret Reserve Prices in eBay Auctions:

Results from a Pokémon Field Experiment. Available on-line at http://eller.arizona.edu/~reiley/papers/SecretReserves.pdf.

Keser, C. (2003). "Experimental games for the design of reputation management systems." IBM Systems Journal **42**(3): 498-506.

List, J. A. and D. Lucking-Reiley (2000). "Demand Reduction in Mulitunit Auctions: Evidence from a Sportscard Field Experiment." The American Economic Review **90**(4): 961-972.

List, J. A. and D. Lucking-Reiley (2002). "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign." Journal of Political Economy **110**(8): 215-233.

Livingston, J. (2002). How Valuable is a Good Reputation? A Sample Selection Model of Internet Auctions, University of Maryland. Available on-line at http://www.wam.umd.edu/~kth/reputation1.pdf.

Lucking-Reiley, D. (1999). "Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet." The American Economic Review: 1063-1080.

Lucking-Reiley, D., D. Bryan, N. Prasad and D. Reeves (2000). Pennies from eBay: the Determinants of Price in Online Auctions, Vanderbilt. Available on-line at http://www.vanderbilt.edu/econ/reiley/papers/PenniesFromEBay.pdf.

McDonald, C. G. and V. C. Slawson Jr. (2002). "Reputation in an Internet Auction Market." Economic Inquiry **40**(4): 533-650.

Melnik, M. I. and J. Alm (2002). "Does a Seller's Reputation Matter? Evidence from eBay Auctions." Journal of Industrial Economics **50**(September): 337-350.

Melnik, M. I. and J. Alm (2003). Reputation, Information Signals, and Willingness to Pay for Heterogeneous Goods in Online Auctions. Available on-line at http://www.gsu.edu/~wwwsps/publications/2003/030101onlineauctions.htm.

Miller, N., P. Resnick and R. Zeckhauser (2005). "Eliciting Honest Feedback in Electronic Markets." Management Science. **51** (9): 1359-1373.

Reiley, D. (forthcoming). "Field Experiments on the Effects of Reserve Prices in Auctions: More Magic on the Internet." RAND Journal of Economics.

Resnick, P. and R. Zeckhauser (2002). Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System. The Economics of the Internet and E-Commerce. M. R. Baye. Amsterdam, Elsevier Science. **11:** 127-157.

Schemo, D. J. (2004). In Online Auctions, Misspelling in Ads Often Spells Cash. New York Times.

Shapiro, C. (1983). "Premiums for High Quality Products as Returns to Reputations." Quarterly Journal of Economics **98**(4): 659-680.

Slonim, R. and A. Roth (1998). "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic." Econometrica **66**(3): 569-596.

Tversky, A. and D. Kahneman (1974). "Under Uncertainty: Heuristics and Biases." Science **185**: 1124-1131.

Yin, P.-L. (2002). Information Dispersion and Auction Prices, Stanford University. Available on-line at http://www.stanford.edu/~pyin/info.pdf.

VII Affiliations

Resnick, School of Information, University of Michigan
Zeckhauser, Kennedy School, Harvard University and Harvard Business School
Swanson, Johnninaswanson, eBay, and Mission Viejo, California
Lockwood, Northwestern University

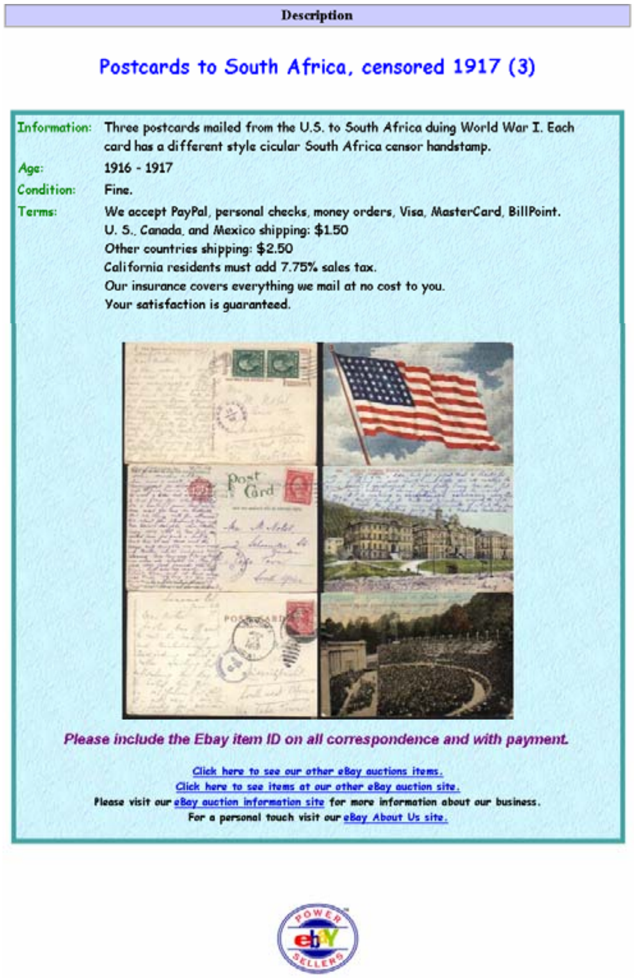
Table 1. Summary of related studies testing impact of reputation on price and probability of sale

Initials	Citation	Items Sold	Mean price	Remarks	Results
HW	(Houser and Wooders forthcoming)	Pentium chips	$244	Sample of items with at least two bids	Positive feedback increases price; negative feedback reduces it
LBPD	(Lucking-Reiley, Bryan, Prasad et al. 2000)	Coins	$173	Censored normal regression	No effect from positive feedback; negative feedback reduces willingness to pay
E	(Eaton 2002)	Electric guitars	$1621	Logit on pr(sale); OLS on price for sold items	No effect from positive feedback. Negative feedback reduces probability of sale only for sellers with >20 feedback; negatve feedback increases price of sold items
BP	(Ba and Pavlou 2002)	Music, software, electronics	$15-2000	Lab experiment in the field: subjects responded with trust level and willingness to pay for auction listings with different feedback profiles spliced in.	Buyers trust more, and pay more if feedback profile has more positives, fewer negatives. Effect is larger for higher priced items
BH	(Bajari and Hortacsu 2003)	Coins	$47	Structural model of bidding based on common values, endogenous entry	Both positive and negative feedback affect probability of modeled buyer entry into the auction, but only positive feedback had a significant effect on final price.
KM	(Kalyanam and McIntyre 2001)	Palm Pilot PDAs	$238	Sample of sold items only	Positive feedback increases price; negative feedback reduces price
MS	(McDonald and Slawson Jr. 2002)	Dolls	$208	Sample of sold items only; simultanteous estimation of price and number of bids	Higher net score (positives – negatives) increases price and number of bids; % negative increase number of bids
MA	(Melnik and Alm 2002)	Gold coins	$33	Censored normal regression	Positive feedback increases price; negative feedback decreases price; negative feedabck reduces probability of sale only
MA2	(Melnik and Alm 2003)	Circulated coins	$93	Censored normal regression	Positive feedback increase price; negative feedback decreases price
DH	(Dewan and Hsu 2004)	Collectible stamps	$37	Sample of sold items only	Higher net score increases price
Y	(Yin 2002)	Computers	$359	Sample of sales with at least 2 bidders. OLS and structural model of bidding	OLS-- no significant effect of reputation; structural model-- more positive feedback causes higher prices because it increases the credibility of information provided in the auction, thus reducing dispersion of bidder valuations of the product

JK	(Jin and Kato 2004)	Sports Trading Cards	$166	Probit on pr(sale); OLS on price for sold items	Positive feedback increases probability of sale; negative decreases probability of sale unless card is professionally graded; no significant effects on price
L	(Livingston 2002)	Golf clubs	$409	Simultaneous estimation of pr(sale), price if sold	More positive feedback increases probability of sale, high bid; severely decreasing returns to positive feedback; no significant effect of negative feedback on probability of sale or high bid
DE	(Dewally and Ederington 2006)	Collectible comic books	$357	Censored normal regression and Heckman estimation	Better reputation increases willingness-to-pay, especially when not professionally graded
CH	(Cabral and Hortacsu 2004)	Laptops, coins, Beanie Babies	$15-$900	Sample of sold items only	After removing one outlier seller, no impact of positives or negatives. Percent negatives has an effect only after eBay starts displaying that statistic in 2003.

Table 2: Sales in the primary experiment.

Week	STRONG sales	NEW sales	Listing prices
1	1A	1B	$9.99-$14.99
2	2A	2B	$9.99
3	1B	1A	$9.99-$14.99
4	2B	1B	$9.99
5	3A	3B	$9.99
6	-	-	-
7	3B	3A	$9.99
8	4A	4B	$14.99-$49.99
9	5A	5B	$4.99-$9.99
10	4B	4A	$14.99-$49.99
11	5B	5A	$4.99-$9.99



Description

Postcards to South Africa, censored 1917 (3)

Information:	Three postcards mailed from the U.S. to South Africa duing World War I. Each card has a different style cicular South Africa censor handstamp.
Age:	1916 - 1917
Condition:	Fine.
Terms:	We accept PayPal, personal checks, money orders, Visa, MasterCard, BillPoint. U. S., Canada, and Mexico shipping: $1.50 Other countries shipping: $2.50 California residents must add 7.75% sales tax. Our insurance covers everything we mail at no cost to you. Your satisfaction is guaranteed.

Please include the Ebay item ID on all correspondence and with payment.

Click here to see our other eBay auctions items.
Click here to see items at our other eBay auction site.
Please visit our eBay auction information site for more information about our business.
For a personal touch visit our eBay About Us site.

Figure 1. A sample lot description for the dealer using his usual seller identity.

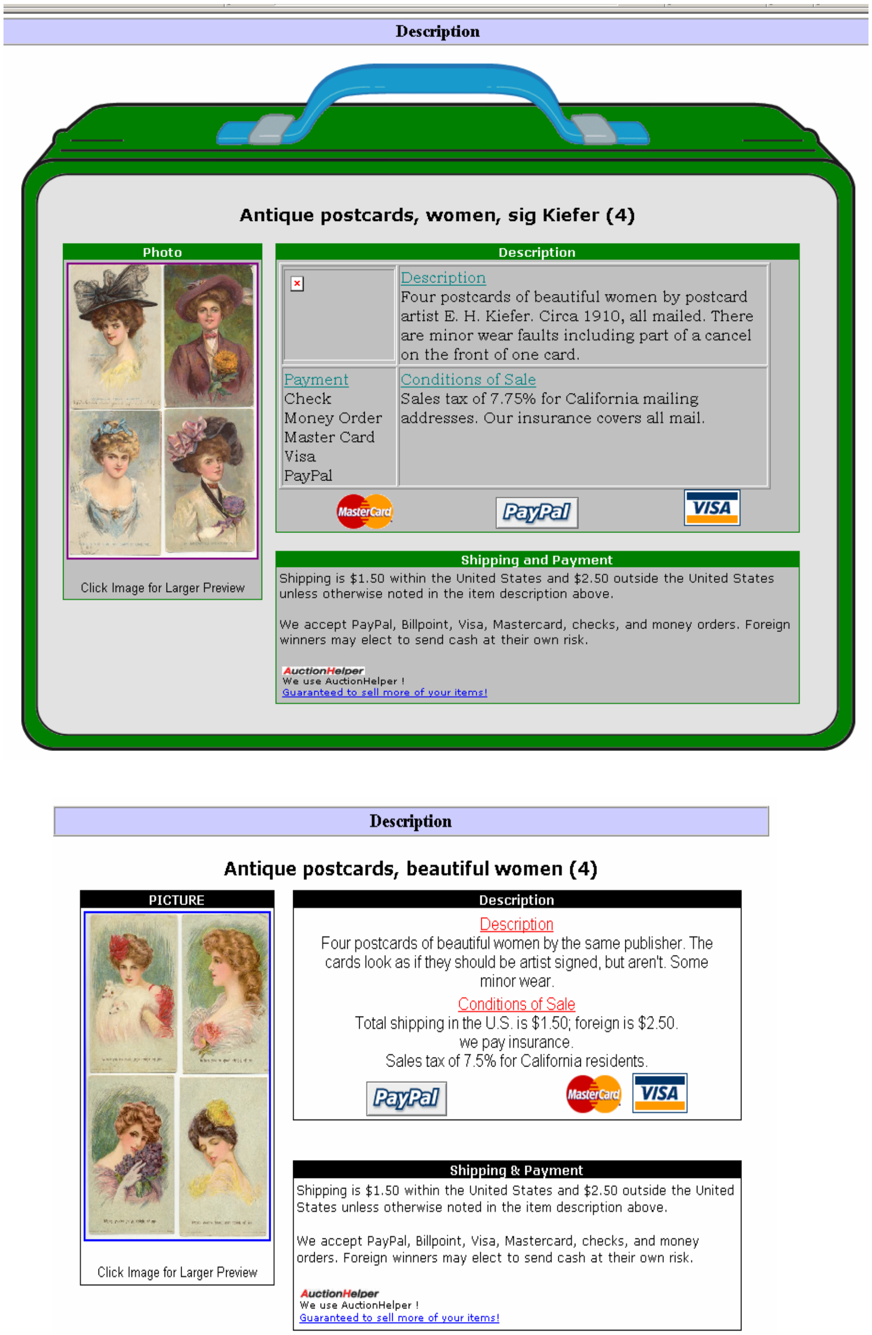


Figure 2. Sample lot descriptions from two of our new sellers, for a matched pair listed in the second experiment. Note that both have the same information available to the buyer, but have different layouts to maintain the impression that they are being listed by different sellers.

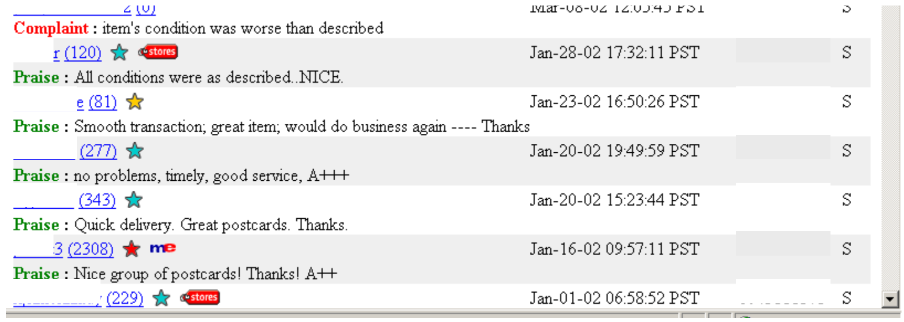


Figure 3. One of the NEW seller’s feedback files.

Table 3: Feedback profiles of new sellers before the start of the second experiment.

Volume in primary experiment	seller with Positives Only	seller with Negatives
High	17+, 0-	12+, 1-
Low	11+, 0-	14+, 2-
Low	9+, 0-	7+, 1-

Table 4. Observed versus expected values of $sign[\ln(STRONG\_price) - \ln(NEW\_price)]$.

Sign	Observed	Expected
Positive	81	69.5
Negative	58	69.5
Zero	59	59
All	198	198

Table 4. Observed versus expected values of
$sign[\ln(STRONG\_price) - \ln(NEW\_price)]$.

Sign	Observed	Expected
Positive	81	69.5
Negative	58	69.5
Zero	59	59
All	198	198

Table 5. Number of sales

	NEW				
STRONG		Not sold	Sold at minimum bid	Sold above minimum bid	Total
	Not sold	47	21	6	74
	Sold at minimum bid	22	12	11	45
	Sold above minimum bid	18	15	46	79
	Total	87	48	63	198

Table 6. Censored normal regression models predicting

$$\ln(STRONG\_price) - \ln(NEW\_price).$$

Const	.078* [-.01, 17]	.091 [-.14, 32]	.110 [-.02, 24]	.048 [-.20,.30]	.106* [-.01, .22]	.017 [-.12, .15]
minimum bid		-.001 [-.02,.02]				
STRONG first			-.060 [-.24,.12]			
NEW2				.021 [-.29, .33]		
NEW3				-.002 [-.38, .38]		
NEW4				.188 [-.12, .5]		
NEW5				-.004 [-.37, .36]		
NEW6				.008 [-.35, .39]		
NEW7				-.197 [-.57,.18]		
Late round					.083 [-.81,.25]	
Pos. feedback						.014 [-.01, .04]
N	151	151	151	151	122	151
LR chi-squared (Pr > chi-squared)		.02 (.90)	.45 (.50)	5.61 (.47)	.99 (.32)	1.38 (.24)

* means P<.05 (one-sided test used for the constant term, late round, and positive feedback, which are expected to have an effect in a particular direction; two-sided test used for the other terms). Ranges in square brackets are 95% confidence intervals.

III

[* *] We gratefully acknowledge financial support from the National Science Foundation under grant number IIS-9977999. Mihir Mahajan provided valuable research assistance. The participants in seminars at the University of Michigan, the University of Arizona, and Carnegie Mellon University provided useful feedback.

[1] Resnick, School of Information, University of Michigan; Zeckhauser, Kennedy School, Harvard University and Harvard Business School; Swanson, Johnninaswanson, eBay, and Mission Viejo, California; Lockwood, Northwestern University.

[2] At eBay and most other on-line auction sites, the norm is for the buyer to send payment first, then for the seller to send the good. Escrow services are available to withhold payment until after acceptance of the shipment, but they are used infrequently.

[3] Researchers have also taken lab experiments, with abstract tasks, to interesting subject pools in field settings (e.g., Slonim and Roth 1998; Henrich, Bowles, Camerer, Fehr, Gintis, and McElreath 2001). Harrison and List (2004) refer to these as *artefactual field experiments*

[4] In 2003, eBay began providing the % of positive scores just after the net feedback rating. At the time of this experiment, only the net feedback rating was shown.

[5] Brick and mortar retailers may be rewarded in part through esteem in the community, but such rewards seem unlikely on eBay. Conceivably such rewards such as self-esteem or adherence to internalized ethics could lead to good behavior despite insufficient direct economic rewards.

[6] The system designer could also impose the penalty; e.g., eBay could charge sellers who develop bad reputations. At present, apart from pursuing fraud situations, eBay imposes no penalties.

[7] If buyers think that sellers can develop reputations as reciprocators, then they could provide positive feedback to get a positive feedback of their own, to bolster their own reputations. However, there is no need for a buyer to have a reputation, unless they too are sellers. EBay merely adds together feedback secured as a buyer and as a seller.

[8] Drawing in part on an earlier version of this paper and this table, Bajari and Hortacsu (2004) also provide a survey of this literature. We, in turn, have benefited from their survey in updating this section of the paper.

[9] Shipping costs, usually paid by the buyer, will affect bidding, especially for inexpensive items, for which shipping costs account for a large proportion of the total expense (Hossain and Morgan, forthcoming). The studies account for these costs in various ways.

[10] If a study truncates the sample by selecting only sold lots or censors the outcome variable by treating the minimum bid as selling price, but utilize an Ordinary Least Squares regression, it will underestimate the effect of a positive factor, such as a strong reputation. The intuition is straightforward. The weaker is this positive factor, the more exceptional will be the required positive error to get included in the sample. Examining only sold lots would thus disproportionately include strong upside outliers when reputation is weak, leading to a biased underestimate of reputation benefits. Studies have dealt with this problem by employing censored normal regression (LBPD, MA, L), or through simultaneous maximum likelihood estimation of both probability of sale and price contingent on sale (L).

[11] In some sense, such controls argue that the seller may not be maximizing, or that the goods sold have other intrinsic differences, e.g., their properties call for different length auctions.

[12] BP secured many of the benefits of a controlled trial by presenting subjects – all real eBay members -- with actual eBay listings spliced together with fake feedback profiles. They then inquired about their

IV

trust in the seller, and the price premium they would pay. This *framed field experiment* in the classification scheme of Harrison and List (2004) allows for comparison across feedback profiles while holding the item presentation constant. However, the subjects knew it was an experiment, might not have been interested in the items presented, and might have made different choices if bidding with their own money.

[13] Two lot pairs were discarded from the analysis because the dealer accidentally listed them with different starting prices for the two sellers. This left a total of 198 pairs for analysis from in the primary experiment.

14 Intuitively, a single negative in a short history is far more predictive of a problem on the next transaction than a single negative in a long history. Outcomes from eBay's transactions of Feb. 20, 1999 , analyzed in RZ, confirm this intuition. Among sales by sellers with 5-20 positive feedbacks, 0.71% (21 of 2948) led to neutral or negative feedback when the seller had no prior neutral or negative feedback, but 2.50% (7 of 280) led to problematic outcomes for those sellers with exactly one previous negative or neutral feedback (chi-sq(1)=9.50, p<.01). Among sales by sellers with 500 or more positive feedbacks, none of the 352 sales by sellers with unblemished records led to neutral feedback, nor did any of the 306 sales by sales with exactly one previous negative or neutral feedback. Whatever difference in risks may exist between these two groups is undetectable in a sample of this size.

Alternatively, we can think of a complete seller history as being like an experiment with the next buyer interpreting the data from that experiment to update her initial beliefs about the next transaction's outcome. Starting from any prior, the Bayes' Factor indicates how far the buyer should update (Bolton, Fong, and Mosquin 2003). The magnitude of the Bayes' factor for the short history including one negative feedback would be much larger than that for the short history without a negative. There is very little absolute difference between the Bayes' factors for long histories with zero or one dissatisfied customer.

[15] We could have conducted the experiment with $.01 minimum bids in order to avoid or at least minimize this problem in analysis, but for ecological validity of the experiment, we preferred to have the dealer follow his normal sales practices as closely as possible.

[16] If the dealer had perfect knowledge, implying perfect price discrimination, all lots would sell at the starting price.

[17] The hypothesized underlying variable is the difference in logs of the market's willingness-to-pay. When STRONG sells the lot but NEW does not, we use the minimum bid as NEW's price and treat the observation as right censored. That is, the market was willing to pay somewhat less than the minimum bid to NEW, so the difference in logs is at least as large as what was observed. Conversely, if NEW sells a lot but not STRONG, the observation is left censored. This approach follows modeling specifications that have been commonly used by other researchers (LBPR, MA).

This modeling specification does not recognize that eBay effectively follows second-price auction rules, i.e., that the market's willingness–to-pay is really the second highest valuation among bidders. When there is more than one bidder, this valuation is observed, but when there is only one bidder, he purchases at exactly the minimum bid, and the second highest valuation is below the observed sale price. If we treat these minimum-bid transactions as censored, they get pooled with lots that did not sell at all, and much less information is available. For example, there were twelve lots that sold at exactly the minimum bid for each of STRONG and NEW. In the result reported in the text, these are treated as equal selling prices. If both are treated instead as left-censored observations, however, there is no information available about the difference between them, and the remaining lots where there is a clear difference have more weight. A censored normal estimation of the difference between ln(STRONG) and ln(NEW) with lots sold at the minimum bid treated as censored rather than actual observations yields a higher estimated mean difference, .178, and the one-side test remains significant (p=.023).

[18] At the median selling price of $14.99, this would correspond to $1.23 per item.

[19] The authors wish to thank Sue Fussell for pointing out this potential confound.

IV

[20] Due to technical difficulties, the full bidding history for both STRONG and NEW was not collected during a three week period in middle of the experiment, which included one week when sales were conducted as part of the experiment.

[21] Only four of the cross-bids were for auctions by the dealer that were not part of the experiment, perhaps reflecting that his other listings were mostly for stamps rather than postcards.