Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Engineering Trust: Reciprocity in the Production of Reputation Information

Gary Bolton, Ben Greiner, Axel Ockenfels,

To cite this article:
Gary Bolton, Ben Greiner, Axel Ockenfels, (2013) Engineering Trust: Reciprocity in the Production of Reputation Information. Management Science 59(2):265-285. http://dx.doi.org/10.1287/mnsc.1120.1609

MANAGEMENT SCIENCE
Vol. 59, No. 2, February 2013, pp. 265–285
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.1120.1609

Engineering Trust: Reciprocity in the Production of Reputation Information

Gary Bolton
Smeal College of Business, Pennsylvania State University, University Park, Pennsylvania 16802; and
Jindal School of Management, University of Texas at Dallas, Richardson, Texas 75080, gbolton@utdallas.edu

Ben Greiner
School of Economics, University of New South Wales, Sydney NSW 2052, Australia, bgreiner@unsw.edu.au

Axel Ockenfels
Department of Economics, University of Cologne, D-50923 Köln, Germany,
ockenfels@uni-koeln.de

Reciprocity in feedback giving distorts the production and content of reputation information in a market, hampering trust and trade efficiency. Guided by feedback patterns observed on eBay and other platforms, we run laboratory experiments to investigate how reciprocity can be managed by changes in the way feedback information flows through the system, leading to more accurate reputation information, more trust, and more efficient trade. We discuss the implications for theory building and for managing the redesign of market trust systems.

Key words: market design; reputation; trust; reciprocity; eBay
History: Received January 28, 2011; accepted May 18, 2012, by Teck Ho, decision analysis. Published online in *Articles in Advance* December 10, 2012.

1. Introduction

This paper reports on the repair of an Internet market trust mechanism. All markets require some minimum amount of trust (Akerlof 1970), but it is a particular challenge for Internet markets, where trades are typically anonymous, geographically dispersed, and executed sequentially. To incentivize trustworthiness, Internet markets often employ a reputation-based "feedback system," enabling traders to publicly post information about past transaction partners. Online markets using a feedback system include eBay, Amazon, and RentACoder, among many others. For these markets, feedback systems with their large databases of transaction histories are a core asset, crucial for user loyalty and market efficiency.

Based on new data and reports from other researchers, we see that the feedback information given in the eBay marketplace exhibits a strong reciprocal pattern (§2). This was a problem because the reciprocity tended to reduce the informativeness of the feedback given and likely hampered market efficiency. We also report on our approach to solving this problem, which combines behavioral economics with an engineering perspective. An engineering study puts the behaviorial science to a *prescriptive* test. Basic theory implies that a reputation system that elicits accurate and complete feedback information can promote trust and cooperation among selfish traders even in such adverse environments as online market platforms (e.g., Wilson 1985, Milgrom et al. 1990). So there is theoretical reason to believe that a properly designed feedback system can effectively facilitate trade. At the same time, the engineering takes us further down the causation chain than present theory goes, to gaming in the *production* of reputation information. Reputation builders retaliate for negative reviews, thereby inhibiting the provision of negative reviews in the first place. The resulting bias in reputation information then works its way up the chain, ultimately diminishing market efficiency. This complication challenges the usefulness of the existing concepts of reputation building that abstract away from the endogeneity of feedback production. One of the major advantages of engineering studies is to identify such gaps in existing concepts and to suggest new research questions (see Ostrom 1990 and Roth 2002 for pioneering work along these lines; Milgrom 2004, Roth 2008, and Greiner et al. 2012 for matching and auction market design surveys; and Chen et al. 2010 for an intriguing design study of social information flows in an online public good environment).

An engineering study is also a method for vetting how the scientifically developed ideas will affect the marketplace *prior* to implementation, to reduce the risk to the marketplace of costly mistakes caused by unforeseen or underestimated circumstances. In our

case, it turned out that the retaliatory behavior on eBay (and other marketplaces) has an institutional trigger in the rules governing feedback timing and observability. Redesigning the feedback system to fix the problem presented three kinds of risks. First, it was not clear how responsive the larger market would be to the fix: To be economically effective, the new system needed to evoke strategically motivated changes in the economic and social behaviors of the traders, regarding both feedback provision and trade conduct, as the information flows through the market. Second, changing the feedback rules risked undesirable side effects. As we will see, reciprocity appears important to getting (legitimately) satisfactory trades reported; eliminating all opportunities for reciprocity (as some redesigns would do) risked a lurch from underreporting negative outcomes to overreporting them. Third, a successful redesign needed to deal with various path dependencies. eBay's feedback system is synchronized with other parts of the market platform, such as eBay's conflict resolution system, so that significant changes in one part would often entail major changes in other parts. Risk is inherent to market redesign generally, so solutions entailing small changes are typically preferred to solutions entailing large changes (Niederle and Roth 2005). The two competing redesigns reflect this principle in that both build on, rather than abandon, the existing system. The *Blind* feedback proposal changes the timing of feedback disclosure, such that one trader's feedback cannot be conditioned on the other's. The detailed seller rating system, which eBay eventually adopted, allows buyers to submit additional, one-sided feedback that is not subject to feedback retaliation. Each proposed system has potential advantages and disadvantages (§2).

Descriptive data from other Internet markets that have feedback systems with features similar to those proposed answer some of our questions (§3), but not all of them: Behavioral and institutional differences across the markets create substantial ambiguity; one proposal, in particular, has major features not shared with any existing market. Also, we lack field data on the underlying cost and preference parameters in the markets and so cannot easily measure how feedback systems affect market efficiency. To narrow the uncertainty, we crafted a test bed experiment designed to capture the theoretically relevant aspects of behavior and institutional changes (§4). In combination with the field observations, the lab data provide a robust picture of how the proposed fixes can be expected to influence feedback behavior and the larger market system. Our analysis guided eBay in its decision to change the reputation system, which allows us to present preliminary data on how the implemented new field system performs (§5). The lessons learned in this study appear to extend beyond the scope of eBay's feedback system, because the reputation-building mechanisms in many markets and social environments, both online and offline, are vulnerable to feedback retaliation (e.g., financial rating services, employee job assessments, word-of-mouth about colleagues). We discuss the implications for theory building about these mechanisms and for managing the design of market trust systems (§6).

2. The Feedback Problem and Two Proposals to Fix It

We first review eBay's conventional feedback system (§2.1). We examine evidence, from new data as well as from the work of other researchers, for a reciprocal pattern in feedback giving and for the role of the rules that govern feedback giving (§2.2). An important point will be that reciprocal behavior appears to have good as well as bad consequences for the system.[1] We then discuss two proposals put forward to mitigate the bad consequences (§2.3).

2.1. eBay's Conventional Feedback System

eBay facilitates trade in the form of auctions and posted offers in more than 30 countries. In 2007, when we collected our data, 84 million users bought or sold $60 billion in goods on eBay platforms. After each eBay transaction, both the buyer and the seller are invited to give feedback on each other. Until spring 2007 (when the system changed), only "conventional" feedback could be left. Under this system, traders could rate a transaction as positive, neutral, or negative (along with a short text comment). Submitted feedback was immediately posted and available to all traders. Conventional feedback ratings could be

[1] That said, many (but not all) studies find that feedback has positive value for the market, as indicated by positive correlations between the feedback score of a seller and the revenue and the probability of sale. See, for example, Bajari and Hortaçsu (2003, 2004), Ba and Pavlou (2002), Cabral and Hortaçsu (2010), Dellarocas (2004), Dewan and Hsu (2001), Eaton (2007), Ederington and Dewally (2006), Houser and Wooders (2005), Jin and Kato (2006), Kalyanam and McIntyre (2001), Livingston (2005), Livingston and Evans (2004), Lucking-Reiley et al. (2007), McDonald and Slawson (2002), Melnik and Alm (2002), Ockenfels (2003), Resnick and Zeckhauser (2002), and Resnick et al. (2006). See Ba and Pavlou (2002), Bolton et al. (2004, 2005), and Bolton and Ockenfels (2009) for laboratory evidence. Further related experimental evidence is provided by Dulleck et al. (2011), who investigated potentially efficiency-enhancing mechanisms in large experimental credence goods markets, which are—like eBay—characterized by asymmetric information between sellers and consumers, and by Sutter et al. (2010), who found large and positive effects on cooperation in an experimental public goods game if group members can endogenously determine its institutional design. Lewis (2011) studied endogenous product disclosure choices of sellers of used cars on eBay as a complementary mechanism to overcoming problems of asymmetric information in the market place.

removed from the site only by court ruling, or if the buyer did not pay, or if both transaction partners mutually agreed to withdrawal.[2]

The most common summary measure of an eBay trader's feedback history is the *feedback score*, equal to the difference between the number of positive and negative comments from unique eBay traders (neutral scores are ignored). A trader's feedback score appears on the site. An important advantage of this score is that it incorporates a reliability measure (experience) in the measure of trustworthiness. The feedback score is also the most commonly used measure of feedback history in research analyses of eBay data.[3]

Observe that the feedback score makes no distinction between feedback for a buyer and feedback for a seller, giving each equal weight in the aggregation. Many individual feedback scores reflect a mix of seller and buyer feedback. In eBay data set 1 (we collected a number of data sets as part of this project; each is described in Online Appendix B, available at http://lboe.utdallas.edu/garyebolton/appendices/), about 65% of the traders were both buyers and sellers at least once, and 50% have completed five or more transactions in both roles.

A second important observation is that most *moral hazard* worries (the opportunities for violating trust) are on the seller side of the market. The buyer renders payment before the seller ships the good. If the buyer fails to send payment, as he was trusted to do, the seller incurs time costs and probably loses the transaction fee, but still has the good for later sale. In contrast, the buyer has to trust that the seller will ship the good and in a timely manner, that the seller's description of the good was accurate, and that the seller will refund or make good if there are problems.[4]

2.2. Reciprocal Feedback, Benefits, and Costs

Feedback information is largely a public good, helping all traders to manage the risks involved in trusting unknown transaction partners. Yet in our data, about 70% of the traders—sellers and buyers alike—leave feedback (a number consistent with previous research).[5] In the following, the null hypothesis is always that feedback is given independently, whereas the alternative hypothesis states that feedback is given conditionally, following a reciprocal pattern. The analysis is based on 700,000 completed eBay transactions taken from seven countries and six categories in 2006/2007.[6]

2.2.1. Feedback Giving. If feedback were given independently among trading partners, one would expect the percentage of time both partners give feedback to be $70\% \times 70\% = 49\%$. Yet mutual feedback is given much more often, about 64% of the time. The top rows of Table 1 contain two related observations: First, both buyers and sellers are more likely to provide feedback when the transaction partner has given feedback first. Second, the effect is stronger for sellers than for buyers; when a buyer gives feedback, the seller leaves feedback 87.4% of the time, versus 51.4% when the buyer does not leave feedback (in a moment we will see that sellers sometimes have an incentive to wait).

[2] eBay's old feedback system was the product of an 11-year evolutionary process. In its first version, introduced in 1996, feedback was not bound to mutual transactions: every community member could give an opinion about every other community member. In 1999/2000 the ability to submit non transaction-related feedback was removed. The percentage of positive feedback as a published aggregate statistic was introduced in 2003, and in 2004 the procedure of mutual feedback withdrawal was added. Since 2005, feedback submitted by eBay users leaving the platform shortly thereafter or not participating in "issue resolution processes" is made ineffective, and members who want to leave neutral or negative feedback must go through a tutorial before being able to do so. In spring 2007 a new system was introduced, as described in §5. In 2008, new features were implemented.

[3] Another common measure is the "percentage positive," equal to the share of positive and negative feedbacks that is positive. For our data, which measure is used makes little difference; we mostly report results using the feedback score.

[4] The text presents a somewhat simplified account of the buyer moral hazard issue. We gathered some anecdotal evidence for buyer moral hazard from our surveys with eBay traders conducted jointly with eBay, from eBay's online feedback forum and from eBay seller conferences. There are four themes: (i) The buyer purchases the item but never sends the payment, as noted in the text. (ii) The buyer has unsubstantiated complains about the item. (iii) The buyer blackmails the seller regarding feedback. (iv) After two months the buyer asks the credit card provider to retrieve the payment (eBay's payment service PayPal does not provide support in these cases). However, beyond anecdotal cases along these lines, buyer moral hazard appeared not to be the critical challenge for eBay and eBay users that seller moral hazard is.

[5] The number varies somewhat across categories and countries. Resnick and Zeckhauser (2002) found that buyers gave feedback in 51.7% of the cases and sellers in 60.6%. Cabral and Hortaçsu (2010) reported a feedback frequency from buyer to seller in 2002/2003 of 40.7% in 1,053 auctions of coins, notebooks, and Beanie Babies. In their 2002 data set of 51,062 completed rare coin auctions on eBay, Dellarocas and Wood (2008) observed feedback frequencies of 67.8% for buyers and 77.5% for sellers.

[6] Online Appendix B contains a list of the field data sets used in this paper. In our description of the field data that motivate our experiment, here as well as in §5, we report mostly descriptives and simple correlations rather than more in-depth regression analysis. We believe that, given the number of observations and the economic size of the reported effects, such "eyeball tests" combined with the cited evidence from other studies will be sufficient to convince readers that reciprocity is an issue. Moreover, our laboratory study provides complementary and highly controlled evidence for these phenomena. Although not reported here, regressions of feedback behavior (e.g., feedback probability, timing, and content) on observables, controlling for various factors such as country and product category, do confirm our findings (see Ariely et al. 2005 and Kagel and Roth 2000 for a similar approach of complementing field with laboratory data).

Table 1 Feedback Giving and Content, Conditional Probabilities, and Correlations

Feedback-giving probability	Partner did not yet give feedback (%)	Partner gave feedback already (%)
Buyer	68.4	74.1
Seller	51.4	87.4

Kendall's tau correlations between seller's and buyer's feedback

	Feedback-content correlation							
	All cases		Buyer gave feedback second		Seller gave feedback second		Feedback-giving correlation	
Country	*N*	Tau	*N*	Tau	*N*	Tau	*N*	Tau
All	458,249	0.710	139,772	0.348	318,477	0.884	725,735	0.693
Australia	20,928	0.746	6,040	0.340	14,888	0.928	31,990	0.752
Belgium	8,474	0.724	3,097	0.464	5,377	0.880	12,301	0.684
France	24,933	0.727	8,095	0.423	16,838	0.883	39,104	0.703
Germany	133,957	0.656	45,836	0.331	88,121	0.840	192,565	0.644
Poland	457	1.000	172	—	285	1.000	1,134	0.783
United Kingdom	93,266	0.694	31,316	0.379	61,950	0.875	143,877	0.692
United States	176,009	0.746	45,133	0.313	130,876	0.911	302,213	0.701

Notes. Observations where feedback was eventually withdrawn are not included in correlations. In the cell with "—," the standard deviation is zero. All other correlations are highly significant.

2.2.2. Feedback Content. Also observe from Table 1 that there is a high positive correlation between the content of buyer and seller feedback within each country sampled. There are likely a number of reasons for this; for example, a problematic transaction might leave both sides dissatisfied. But Table 1 also provides a first hint that reciprocity in feedback content has a strategic element: If feedback were given independently, the correlation between seller and buyer content, as measured by tau, should be about the same when the seller gives feedback second as when the seller gives feedback first. In fact, the correlation is about twice as high when the seller gives feedback second. The pattern is similar across countries.

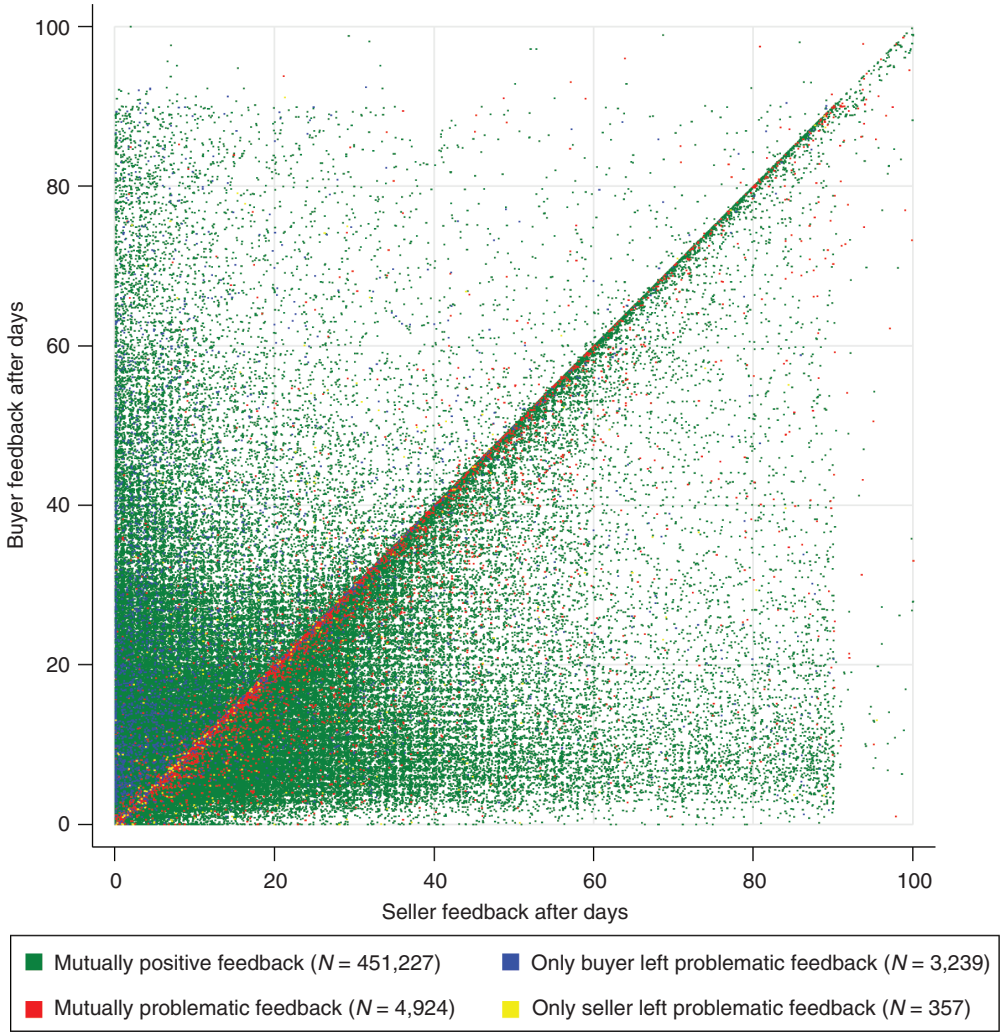
2.2.3. Feedback Timing. If feedback timing were independent among trading partners, one would expect the timing of buyer and seller feedback to be uncorrelated with content. But this is not the case: Figure 1 shows the distribution of feedback timing for those transactions where both traders actually left feedback. The green dots represent the timing of mutually positive feedback. More than 70% of all these observations are located below the 45-degree line, indicating that in most cases, the seller gives feedback after the buyer. The red dots visualize observations of mutually problematic feedback. Here the sellers' feedback is given second in more than 85% of the cases. Moreover, mutually reciprocal feedback is much more heavily clustered alongside the 45-degree line than nonreciprocal feedback. For instance, a seller who gives negative feedback does so much faster after the buyer gave negative feedback than after the buyer gave positive feedback; the median number of days since the buyer gave negative feedback (standard deviation) is 0.77 (11.1), or 2.98 (17.9) if the buyer gave positive feedback. All these differences in timing are significant at all conventional levels.

The tightness and sequence in timing suggest that sellers reciprocate positive feedback and "retaliate" for negative feedback. Seller retaliation also explains why more than 70% of cases in which the buyer gives problematic feedback and the seller gives positive feedback (blue dots in Figure 1) involve the buyer giving feedback second—the buyer going first would involve a high risk of retaliation. Observations in which only the seller gives problematic feedback (yellow dots) are rare and have their mass below the 45-degree line.

Why do sellers retaliate for negative feedback? Existing theory and laboratory studies on reputation building, although not developed in the context of the production of reputation information, suggest multiple strategic and social motives (and these dovetail well with anecdotal and survey evidence that we have collected).[7] Some retaliation is probably driven by social preferences or emotional arousal: The buyer's negative feedback harms the seller's reputation, and this triggers the buyer to respond in kind. Retaliating for negative feedback may also help deter negative feedback in the future, because retaliation is viewable by buyers in a seller's feedback history.

[7] See, for example, Kreps and Wilson (1982), Milgrom et al. (1990), Greif (1989), Camerer and Weigelt (1988), Neral and Ochs (1992), Brandts and Figueras (2003), and Bolton et al. (2004) for the strategic role in reciprocity, and see Fehr and Gächter (2000) and the surveys in Cooper and Kagel (2013) and Camerer (2003) for the social aspect in reciprocity. Herrmann et al. (2008) provide cross-cultural evidence for antisocial reciprocity in laboratory experiments where high contributors to public goods are punished by low contributors.

Figure 1 Content and Timing of Mutual Feedback on eBay



Notes. The scatter plot reports about 460,000 observations where both transaction partners gave feedback. "Problematic" feedback includes negative, neutral, and withdrawn feedback.

Also, giving negative feedback increases the probability that the opponent will agree to mutually withdraw the feedback.

2.2.4. Benefits and Cost of Reciprocal Feedback. The main benefit of reciprocal feedback, for both the individual traders involved and the larger system, is that it helps record mutually beneficial trades. A common buying experience on eBay, after a transaction has gone smoothly, is to receive a note from the seller saying he gave you positive feedback and asking you to provide feedback, or saying that he would give you feedback once you left feedback on him (playing or initiating a kind of "trust game"). The data (top of Table 1) suggest that this is an effective tactic for reputation building. It is good for the system too, because mutually satisfactory trading experiences get recorded.

However, in the form of seller retaliation, reciprocal feedback imposes costs both on the buyers retaliated against and potentially on the larger system. With regard to buyers, it hurts them in future trading circumstances where there might be buyer moral hazard (although as noted, this is not frequent). But also recall that many buyers become sellers (§2.1), so negative feedback can hurt them in that role, too. Also, buyers (as with any eBay member) seem to put a high value on their profile, for reasons that cannot be fully explained with only strategic motives (Ockenfels and Resnick 2012). With regard to the larger system, the worry is that it has a chilling effect on buyers' reporting bad experiences out of fear that it will be retaliated. This would bias feedback information to be overly positive and therefore less informative in identifying problem sellers. The fact that from 742,829 eBay users (data set 1; see Online Appendix B) who received at least one bit of feedback, 67% have a percentage positive of 100%, and 80.5% have a percentage positive of greater than 99%, provides suggestive

support for the bias. The observation is in line with Dellarocas and Wood (2008), who examined the information hidden in the cases where feedback is not given. Under some auxiliary assumptions, they estimated that buyers are at least mildly dissatisfied in about 21% of all eBay transactions, a percentage far higher than the levels suggested by the reported feedback. Dellarocas and Wood (2008) argued that many buyers do not submit feedback at all because of the potential risk of retaliation.

Other studies provide complementary evidence on the social and strategic aspects of feedback production. Resnick et al. (2000) and Resnick and Zeckhauser (2002) observed strong correlations between buyer and seller feedback in their eBay field data. The analysis above replicates this finding. Regarding feedback giving, Bolton and Ockenfels (2011) reported a controlled field experiment conducted on eBay with experienced eBay traders. They found that sellers who did not share the gains from trade in an equitable manner received significantly less feedback than sellers who shared equitably. This finding lends additional credence to the suspicion that fear of retaliation is a factor behind dissatisfied buyers staying silent. On a more general level, there is evidence for a common and strong tendency for lenient and compressed performance ratings, as discussed, for instance, in the literature on the "leniency bias" and "centrality bias" in human resource management (Bretz et al. 1992, Prendergast and Topel 1993, Prendergast 1999). Regarding feedback timing, Jian et al. (2010) confirmed that eBay buyers and sellers often employ a conditional strategy of giving feedback. Exploiting information about the timing of feedback provision when the partner does not provide feedback, Jian et al. (2010) estimated that, under auxiliary assumptions, feedback is conditional 20%–23% of the time. Ockenfels and Resnick (2012) provide a more extensive survey of the literature. Overall, this literature, based on a variety of field data sets, is consistent with the patterns of social and strategic feedback usage that we find in our data and provide the starting point of our engineering approach.

2.3. Two Alternative Redesign Proposals

Any institutional change in a running market must respect certain path dependencies. This is particularly true for reputation systems, which by their nature connect the past with the future. For this reason, the redesign proposals we consider carry forward (in some form) the conventional ratings of the existing system, allowing traders to basically maintain the reputation they built before the change.[8] At the same time, each proposal attacks one or the other of two features that appear to facilitate retaliation behavior, either the open, sequential posting that allows a trading partner to react to the feedback information or the two-way nature of the ratings that allows sellers to retaliate against buyers.[9]

Proposal 1: *Make conventional feedback double blind*. That is, conventional feedback would only be revealed after both traders submitted feedback or after the deadline for feedback submission expired. Thus, a trader could not condition her feedback on the feedback of her transaction partner's, thereby excluding sequential reciprocity and strategic timing and making seller retaliation more difficult. The conjecture is that this will lead to more accurate feedback. A double-blind system of this sort has been suggested by Güth et al. (2007), Reichling (2004) and Klein et al. (2007), among others. A major risk with a double-blind system concerns whether it will diminish the frequency of feedback giving, particularly with regard to mutually satisfactory transactions. Because trading partners effectively give feedback simultaneously, giving positive feedback could not be used to induce a trading partner to do the same. Another issue is that a seller can game the system by preventing the publication of received feedback, potentially of value to other traders, until the end of the feedback deadline by not submitting feedback herself.

Proposal 2: *Supplement the existing conventional feedback system with a one-sided feedback option that enables buyers to give a detailed seller rating* (*DSR*). In principle, a one-sided system in which only the buyer gives feedback is the surest way to end seller retaliation. Such a system has been proposed by Chwelos and Dhar (2007), among others. But although there is more scope for moral hazard on the seller side than on the buyer side in eBay's marketplace, there might be room for buyer moral hazard as well. Moreover, gaining positive feedback as a buyer appears to be an important step for many traders in their transition to a successful seller. For these reasons, the proposal was to create a DSR system to supplement the conventional feedback system: Conventional feedback would

[8] Another example for the consideration of path dependency in practical reputation system design can be found on Amazon.com. When changing its ranking of voluntary book reviewers in 2008, Amazon retained its classical system (tracking lifetime quantity of reviews) while adding new measures to reflect the quality of reviews.

[9] Other options were considered in the process of developing "Feedback 2.0" but were discarded relatively quickly in favor of the two explored here. Most notably, we considered a system that has feedback given only by buyers or strictly separates feedback earned as a seller and feedback earned as a buyer. Miller et al. (2005) proposed a scoring system that makes reporting honest feedback in the absence of other feedback-distorting incentives part of a strict Nash equilibrium, but they did not consider the problem of reciprocally biased feedback.

Table 2 Feedback Frequency, Content, and Correlation on MercadoLivre and eBay China, Compared to Other eBay Platforms

	Feedback frequency			Problematic feedback given by (%)		Feedback-content correlation	Feedback-giving correlation
	N	Buyer (%)	Seller (%)	Buyer	Seller	Kendall's tau	Kendall's tau
eBay U.S.	10,169	74.8	76.7	1.4	1.2	0.720	0.595
eBay Germany	14,297	77.3	76.9	1.9	1.1	0.621	0.623
eBay China	2,011	9.3	19.7	5.0	6.7	0.576	0.652
Verified buyers	1,062	15.0	13.6	5.0	4.9	0.576	0.682
Unverified buyers	949	3.1	3.6		14.7		0.460
MercadoLivre Brazil	1,958	71.1	87.9	18.7	29.2	0.785	0.175

Note. All correlations are highly significant.

be published immediately as usual, but (only) the buyer would have the option to leave additional feedback, blind to the seller.[10] A possible negative consequence is that the conventional and DSR feedback given to sellers might diverge, with unhappy buyers giving positive conventional feedback to avoid seller retaliation, and then being truthful with the (blind) DSR score. This might not be a problem for experienced traders who would know to pay exclusive attention to DSR scores. But it might make it harder and more costly for new eBay traders to learn how to interpret reputation profiles. For some traders, the inconsistency might damage the institutional credibility of the feedback system.

3. Descriptive Evidence from Other Internet Markets

As a first step in evaluating the two proposals, we searched for and examined systems involving double-blind and one-sided feedback in other Internet markets. The benefit of field data is that we can study behavior in naturally evolved environments. At the same time, there are limitations to the conclusions we can draw. We first review the data, then discuss the limitations.

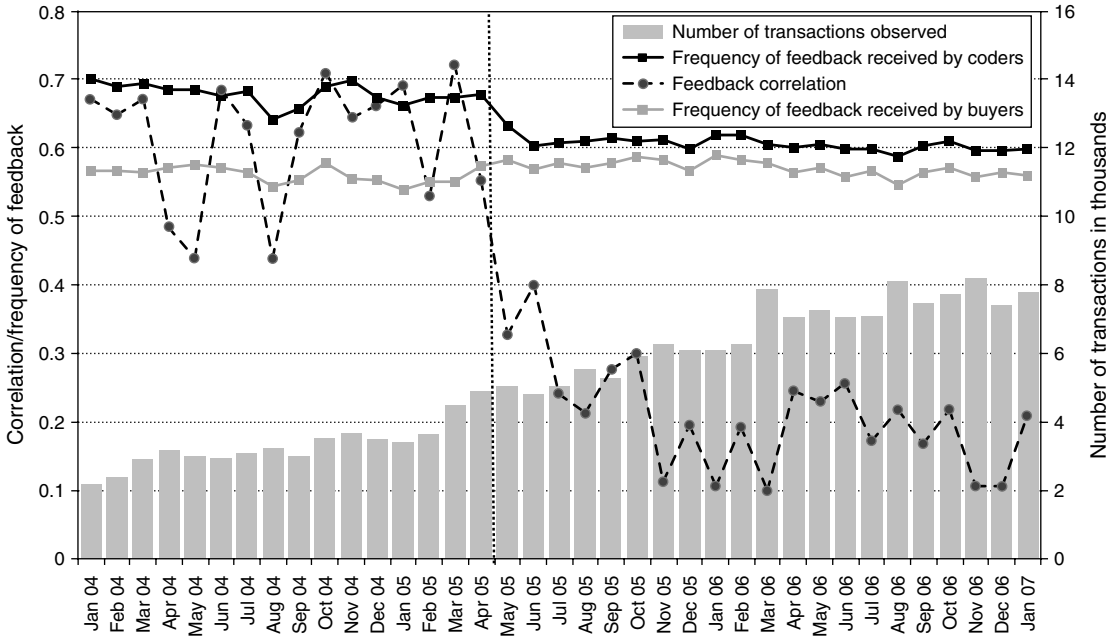
We start with data culled from two markets with double-blind systems similar to the Proposal 1 system (§2.3). The first field evidence comes from eBay's own market in Brazil. MercadoLivre began in 1999 as an independent market, eBay-like in its objective but with some unique trading procedures. eBay bought the market in 2001 and decided to keep some elements, including a double-blind feedback system. MercadoLivre reveals submitted feedback after a 21-day *blind period* that starts on completion of the transaction. No feedback can be given after the blind period has lapsed.

Table 2 shows feedback statistics based on a total of 24,435 completed transactions in data set 3 (see Online Appendix B), which was specifically compiled to compare feedback behavior in eBay's conventional feedback system to other eBay sites (the verified buyer breakout for eBay China will be discussed later in this section). Observe that the share of problematic (negative, neutral, and withdrawn) feedback given on MercadoLivre is multiple times higher than on other mature eBay platforms that do not employ a blind feedback system. Moreover, although the correlation of feedback content differs little from that in other markets (column 7 in Table 2), the correlation of feedback giving is much lower in Brazil than in the United States, Germany, or China (column 8 in Table 2). That is, in those cases where both transaction partners leave feedback, the content in Brazil is as correlated as in the other countries, but the probability of two-way feedback giving is much smaller. One worry we raised with a double-blind system is that diminishing reciprocal opportunities might diminish the rate at which traders leave feedback. But MercadoLivre provides no evidence that double-blind feedback decreases participation; the feedback frequency of 71% for buyers is in line with what we observe in other countries, and 88% of sellers provides even more feedback.

The RentACoder.com site enables software coders to bid for contracts offered by software buyers. RentACoder.com used to have a two-sided, open feedback system, similar to eBay, but switched to a double-blind system in May 2005. RentACoder's motive for the switch (as stated on its help page) is the potential threat of retaliatory feedback in an open system. The double-blind system allows buyers and coders to leave feedback on one another within a period of two weeks after completion of a project.

[10] Another advantage is that we can fine-tune the scaling of the new ratings without disrupting the three-point conventional ratings; the latter would create a number of path dependency problems. Research in psychology suggests that Likert scaling of five or seven points is optimal (e.g., Nunnally 1978, and more recently, Muniz et al. 2005). Additionally, several studies have found that users generally prefer to rate on more categories than submit just one general rating (e.g., Oppenheim 2000). We describe the economic effects of scaling in §4.4. The specific method for posting detailed seller ratings is best understood in the context of a number of practical considerations and is described at the beginning of §5.

Figure 2 **Feedback Frequency and Correlations Before and After the System Change in April 2005 on RentACoder.com**



The RentACoder.com panel data (data set 4, see Online Appendix B) comprises 192,392 transactions. Unlike MercadoLivre, it allows for a within-site comparison, keeping all institutions but the feedback system fixed, and allowing an analysis of the transition from an open to a double-blind system. The transition has no significant effect on average feedback content received by either buyers or sellers, although there is a weakly significant, small increase in the standard deviation of feedback that buyers received.[11] There are, however, other effects indicative of diminishing reciprocity. First, as shown in Figure 2, the monthly correlation between feedback content drops sharply and significantly from an average of 0.62 in the 15 months before the change to 0.21 in the 21 months after the change. We also observe from Figure 2 (and backed by time series regressions controlling for trends) that coders get significantly less feedback after introduction of double-blind feedback, while buyers get a small but significant increase.

The MercadoLivre and RentACoder data are consistent with the claim that a double-blind feedback system leads to buyers giving more discerning feedbacks, with less correlation of feedback between trading partners. The evidence on changes in the frequency of feedback giving is mixed, with the MercadoLivre system showing a high degree of feedback giving, whereas the introduction of the RentACoder double-blind system was followed by a decrease in feedback giving.

[11] Because of space limitations, we omit the regressions of time series of monthly averages on constant, time trend, and blindness dummy, which confirm the observation.

The second set of field evidence comes from markets with *one-way* feedback systems, each similar in some respects to the Proposal 2 system (§2.3). The first evidence comes from a within-platform comparison on the Chinese eBay site, where there is a large proportion of so-called "unverified buyers"—buyers who did not provide proof of their identity (yet). Feedback given by unverified buyers does not count toward the seller's reputation. Thus, from a reciprocity perspective, giving feedback to unverified buyers is much like giving one-sided feedback. Table 2 shows frequency and content of feedback for verified and unverified buyers. We observe that verified buyers receive and give about five times as much feedback as unverified buyers ($\chi^2 = 82.6$, $p < 0.001$) and that feedback giving is much more correlated with verified buyers (the correlation coefficients are 0.460 versus 0.682).[12]

More evidence comes from Amazon.de, which has a one-sided buyer-to-seller feedback system (data set 5, a sample of 320,609 instances of feedback; see Online Appendix B).[13] In addition, we conducted a small email-based survey with a subset of sellers in

[12] Moreover, unverified buyers receive neutral or negative feedback 14.7% of the time in our sample, whereas verified buyers receive negative feedback only 4.9% of the time ($\chi^2 = 2.82$, $p = 0.093$), suggesting that a one-sided system will elicit less positive (and probably more accurate) feedback. However, here, the causality appears to be less clear. Unverified buyers might be more likely to be unfamiliar with the trading and communication norms or to have less long-term interest in the site and so less incentive to build up a good reputation.

[13] Strictly speaking, both sellers and buyers on Amazon are able to submit feedback on each other. However, feedback given to buyers is not accessible to other sellers, whereas feedback to sellers is

our sample. Taking the survey responses of 91 Amazon sellers and the field data together, we find that feedback is left by buyers in about 41% of transactions; if we weight the answers by number of transactions, we get a 36% figure (implying that very active sellers get somewhat less feedback), about half the rate of feedback on the various eBay platforms. We also observe that Amazon feedback exhibits higher variance than does eBay's conventional feedback; only 81.5% of feedback is given in the best category, with a score of 5, whereas middle and low scores of 4, 3, 2, and 1 are given in 14.5%, 2.2%, 1.0%, and 0.9%, respectively, of all cases.

The Chinese eBay and Amazon.de data are consistent with the claim that a one-sided feedback system leads to buyers giving more discerning feedback. At the same time, both markets reinforce the suspicion that removing the opportunity for reciprocal feedback from the system lowers feedback frequency.

Altogether, the field data are suggestive of the potential both proposed fixes to the eBay system have for generating a more accurate, or at least a more dispersed, reflection of trader satisfaction. At the same time, given the highly complex and diverse environments these markets operate in, it is difficult to make clear causal inferences based on the field data alone. For instance, the low level of positive feedback in MercadoLivre may stem from uncontrolled cross-country effects regarding different norms of trading or feedback giving, or from differences in Brazilian payment or postal services (see Özer et al. 2012 for a comparison of Chinese and American trust in information sharing). Similarly, a comparison of RentACoder.com with eBay is complicated by the fact that the RentACoder.com feedback is on a 10-point scale, the market is smaller, the bidding process and price mechanism are different (coders bid for contracts and buyers do not need to select the lowest price offer), etc. With regard to the one-sided proposal, neither the Chinese eBay site nor the Amazon .de site shares the two-way reporting component of the proposed DSR system (in fact, we know of no system with this combined feature).

Along the same lines, and just as important, the field data provide no direct evidence that the reduction in reciprocity improves either the informativeness of feedback or market efficiency. One reason to wonder is that the market in the sample closest to the eBay markets in question, MercadoLivre, exhibits a far higher rate of negative feedback than any other market.[14] Another reason is the relatively low rates of feedback giving in some of the markets with double-blind or one-sided feedback: a substantial drop in feedback giving might raise its own credibility issues, effectively substituting one trust problem for another. With the exception of RentACoder.com, there is little in the way of before-and-after data to guide such an analysis.

published publicly. As a result, sellers typically do not leave feedback. This makes Amazon's system effectively a one-sided system.

[14] One response to this concern is that the rate of negative feedback on MercadoLivre accords well with rates of unhappiness uncovered by research (e.g., Dellarocas and Wood 2008). However, as the experiment reported in the next section makes clear, we should expect more informative feedback to ignite a number of endogenous effects in the system, starting with buyers better identifying and shunning untrustworthy sellers, and so the proportion of unsatisfactory trades should be lower than the present rate of unhappiness.

4. The Laboratory Study

The experiment speaks to the limitations of the field evidence discussed at the end of the last section. Accordingly, the experiment is designed as a level playing field for comparing the performance of the competing feedback system proposals. Experimental controls help us identify the role of reciprocal behavior in the context of feedback giving and establish causal relationships between feedback and market performance (e.g., efficiency). To do these things, the experiment needs to abstract away from a number of features that arise in the natural environments. We will argue that the *combined* laboratory and field data make for a more compelling engineering argument than either kind of data in isolation.

Section 4.1 outlines the experimental design. Section 4.2 shows that the laboratory feedback behavior we observe mirrors key field observations from the conventional system and that different systems lead to different feedback behavior. Section 4.3 measures the impact of the feedback system on the economic performance of the auction market. Section 4.4 shows how market performance is connected to feedback informativeness. Section 4.5 discusses what the combined lab and field data tell us.

4.1. Experimental Design and a Hypothesis

The experiment simulates a market where there is seller moral hazard and includes an auction component that is fixed across all treatments; the feedback component is varied to capture the various scopes for reciprocity across alternative feedback systems.

4.1.1. Auction Component. Each treatment simulates a market that consists of 60 rounds. In each round, participants are matched in groups of four, one seller and three potential buyers. Each buyer i receives a private valuation for the good, v_i, publicly known to be independently drawn from a uniform distribution of integers between 100 and 300 experimental currency units (ECUs). Buyers simultaneously

submit bids of at least 100 ECUs or withdraw from bidding. The bidder with the highest bid (earliest bid in case of a tie) wins the auction and pays a price p equal to the second highest bid plus a 1 ECU increment, or his own bid, whichever is smaller. If there is only one bid, the price is set to the 100 ECU start price. After the auction, all participants in the group are informed of the price and of all bids but the highest.[15] The price is shown to the seller s, who then determines the quality of the good $q_s \in \{0, 0.01, \ldots, 0.99, 1\}$. Permitting quality choice is a simplification of the many potential dimensions of seller moral hazard in the field, like inaccurate item descriptions, long delivery time, low quality, etc. The payoff (not including feedback costs described below) to the seller is $\pi_S = p - 100q_s$ and to the winning buyer i is $\pi_i = q_s v_i - p$.

There were 32 participants in a session and two sessions per treatment. Eight sequences of random parameters (valuations and role and group matching), involving eight participants each, were created in advance. Thus, random group rematching was restricted to pools of eight subjects, yielding four "subsessions" per session and eight statistically independent observations per treatment. To ensure a steady growth of experience and feedback, random role matching was additionally restricted such that each participant became a seller twice every eight rounds. The same eight random game sequences were used in all treatments. Participants were not informed about the matching restriction.

4.1.2. Feedback Component. When the auction ends in a trade, both buyer and seller have the opportunity to give voluntary feedback on the transaction partner. Giving feedback costs the giver 1 ECU, reflecting the small effort cost when submitting feedback. Because our primary interest was long-run effects and not transitional dynamics, we had each subject experience only one feedback system. The underlying assumption here is that there is little in the way of behavioral path dependencies that affect long-run performance.

In the *Baseline* treatment, both the seller and the buyer can submit conventional feedback (CF), rating the transaction as negative, neutral, or positive. Feedback giving ends with a *soft close*: In a first stage, both transaction partners have the opportunity to give feedback. If both or neither gives feedback, then both are informed about the outcome and the feedback stage ends. If only one gives feedback, the other is informed about that feedback and enters the second feedback stage, where he again has the option to give feedback and so a chance to react to the other's feedback.[16] As on eBay, a trader's conventional feedback is aggregated over both buyer and seller roles as the feedback score and the percentage of positive feedbacks (see §2). When the participant becomes a seller, these scores are presented to potential buyers on the auction screen prior to bidding.

The *Blind* treatment differs from the *Baseline* treatment only in that we omit the second feedback stage. That is, buyer and seller give feedback simultaneously, not knowing the other's choice.

The *DSR* (*Detailed Seller Rating*) treatment adds a rating to the *Baseline* treatment feedback system. After giving CF, the buyer (and only the buyer) is asked to rate the statement "The quality was satisfactory" on a five-point Likert scale: "I don't agree at all," "I don't agree," "I am undecided," "I agree," "I agree completely." As in the *Baseline* treatment, we implement a soft close design, but in case the seller delays and enters the second feedback stage, she is only informed about the conventional feedback given by the buyer, not about the detailed quality rating. Number and average of received detailed seller ratings are displayed at the auction page.

All sessions took place in April 2007 in the Cologne Laboratory for Economic Research. Participants were recruited using the online recruitment system ORSEE (Greiner 2004). Overall, 192 students (average age 23.8 years, 49% male) participated in six sessions. After reading instructions (see Online Appendix A, available at http://lboe.utdallas.edu/garyebolton/appendices/) and asking questions, participants took part in two noninteractive practice rounds. Each participant received a starting balance of 1,000 ECUs to cover potential losses. Sessions lasted between one and a half and two hours. At the end of the experiment, the ECU balance was converted to euros at a rate of 200 ECU = 1 euro and was paid out in cash. Participants earned 17.55 euros on average (standard deviation is 2.84), including a show-up fee of 2.50 euros and a 4 euro bonus for filling in a post-experiment questionnaire.

4.1.3. Hypothesis. The experiment has a finite number of trading rounds. Assuming that all agents are commonly known to be selfish and rational, the

[15] Our experimental design, including features such as the handling of increments and the information provided to bidders, is chosen analogously to eBay's rules. However, for simplicity, we chose a sealed-bid format and abstracted away from eBay's bidding dynamics, which is known to create incentives for strategic timing in bidding (Roth and Ockenfels 2002).

[16] This mirrors the feedback strategies admitted on eBay in simplified form. On eBay, there is always a possibility to respond to submitted feedback. So the basic types of strategies a trader can pursue are: do not submit feedback at all, submit unconditional feedback, or submit feedback conditional on other's feedback and otherwise don't submit. Our soft close design captures these strategic options. Ariely et al. (2005) and Ockenfels and Roth (2006) model the ending rule of Amazon.com auctions in a similar way, allowing buyers to always respond to other bids.

unique subgame-perfect equilibrium in all treatments of the experiment stipulates zero feedback giving and quality tendered, with no auction bids. The socially efficient outcome has the bidder with the highest valuation winning the auction, the seller producing 100% quality, with no (costly) feedback giving. So both of these rather extreme scenarios leave no role for the feedback system. If, as seems more likely, feedback is used to build up reputation and to discriminate between sellers, we hypothesize that reciprocal feedback hampers market efficiency because reciprocity compresses reputation scores in a way that makes it harder for buyers to discriminate between sellers; these sellers then have less incentive to deliver good quality.[17] Consequently, the two proposed redesigns, if they diminish the role of reciprocity, should do better.

It is important to note that the experiment focuses on *seller* moral hazard, excluding buyer moral hazard, because each winning bid is automatically transferred to the seller. So feedback given by sellers cannot have grounds in the transaction itself. Three considerations guided us in this design choice. First, on eBay, the scope for buyer moral hazard is relatively small (§2.1). Second, seller retaliation for negative feedback was perceived as a much larger problem for eBay than was buyer retaliation, a perception confirmed by Figure 1, where 85% of mutually negative feedback begins with the buyer going first.[18] Third, not admitting buyer moral hazard removes an important confound in interpreting negative seller feedback. In the experiment, negative feedback given by the seller is clearly retaliatory feedback. Negative seller feedback also imposes a cost on the buyer in the form of potentially adverse affects of the buyer's future profits as a seller, as is the case for a majority of traders on eBay (§2.1).

Table 3 Timing of Feedback

	Baseline (%)	*Blind* (%)	*DSR* (%)
Both first round	27	26	29
None first round	16	24	15
Seller 1st, buyer in 2nd	4		2
Seller 1st, buyer not (in 2nd)	5	8	8
Buyer 1st, seller in 2nd	24		17
Buyer 1st, seller not (in 2nd)	23	42	28

4.2. Feedback Behavior

In this section, we investigate whether the feedback pattern in the *Baseline* treatment mirrors the pattern observable in the field and how the feedback behavior in the alternative systems compares. Unless indicated otherwise, any statistical tests reported here and in subsequent sections are two-tailed Wilcoxon matched pairs signed ranks tests relying on the (paired) fully independent matching group averages.

4.2.1. Feedback Giving. In the *Baseline* treatment, buyers give feedback in about 80% and sellers in about 60% of the cases, with an average of about 70%, just like in the field data. Relative to *Baseline*, *Blind* exhibits significant drops in both buyer (68%) and seller (34%) giving frequencies ($p < 0.025$ in both cases), whereas *DSR* exhibits only minor and insignificant reductions for both buyers (77%) and sellers (57%; $p > 0.640$ in both cases).[19]

4.2.2. Feedback Timing. When possible, sellers are more likely than buyers to wait until the other has given feedback (Table 3; $p < 0.025$ both in *Baseline* and *DSR*). This effect is most pronounced when feedback is mutually neutral/negative; the only case with buyers more often moving second is when the buyer gives problematic and the seller positive conventional feedback (for details, see Table 9 in Online Appendix C, available at http://lboe.utdallas.edu/garyebolton/appendices/). These interaction patterns of feedback content and timing are very similar to what is observed in the field (§2), and thereby reassure us of the suitability of our experimental design of the *CF* feedback component.

4.2.3. Feedback Content. Table 4 shows correlations between conventional feedback across treatments. We find that blindness of feedback significantly decreases the correlation, compared with the open systems. The high correlations in the latter are mainly driven by the cases where sellers delay their

[17] For an overview on different modeling approaches to seller reputation, see Bar-Isaac and Tadelis (2008). There is also an experimental literature testing reputation theory. More recent contributions include Grosskopf and Sarin (2010), allowing for reputation to have either beneficial or harmful effect on the long-run player, and Bolton et al. (2011), searching for information externalities in reputation building in markets with partners and strangers matching, as predicted by sequential equilibrium theory. These as well as other papers (see references cited therein) come to the conclusion that reputation building often interacts with social preferences in subtle ways, often (but not always) making reputation mechanisms more beneficial than predicted by theory, based on selfish behavior. Our study complements this literature by showing how reciprocity can both hamper and promote the effectiveness of reputation mechanisms.

[18] On eBay, there are additional strategic reasons for reciprocity in feedback giving, having to do, say, with building up a reputation of being a "retaliator." Our experiment does not provide the information necessary to employ such complex strategies. The experiment shows that the more direct reciprocal concerns are sufficient to capture much of what we see in the field. To the extent that traders employ more complex reciprocal strategies in the field, our experiment tends to underestimate the effect from feedback reciprocation.

[19] Regression analyses considering interaction effects of treatments with quality support the finding (Table 8 in Online Appendix C) and furthermore show that buyers give feedback significantly more often when quality is low in both alternative designs.

Table 4 Kendall Tau Correlations Between Seller and Buyer Feedback by Timing

	Both 1st	Seller 1st, buyer 2nd	Buyer 1st, seller 2nd	All
Baseline	0.359	0.536	0.901	0.680
Blind				0.411
DSR	0.533	0.730†	0.913	0.759

Note. All correlations highly significant at the 0.1% level, except for the cell indicated by "†," which is weakly significant at the 10% level.

feedback and give feedback second, whereas when both transaction partners give feedback in the first stage, correlations are comparable to blind feedback. However, correlations of simultaneously submitted feedback are significantly different from zero, too.

4.2.4. Negative Feedback. Finally, the probit estimates in Table 5 show the determinants of problematic feedback given to sellers conditional on the buyer giving feedback (where, as before, problematic feedback is defined as either a negative or neutral feedback). Model 1 shows that there is no significant treatment effect overall. But from Model 2, controlling for quality, price, and other factors, we see that problematic conventional feedback increases in both *Blind* and *DSR*. The coefficient estimates for the two treatment dummies are nearly identical, indicating that the size of the effect is about the same in both treatments.[20] The reason for more negative feedback is that buyers receiving poor quality are more likely to give problematic feedback under the alternative systems. More specifically, Figure 3 illustrates that in all treatments, positive conventional feedback (and the highest DSR) is awarded for quality of 100%; likewise, very low quality receives negative feedback in all cases. The major difference between the treatments happens between 40% and 99% quality; here average conventional feedback given is tougher in *Blind* and *DSR*. Also observe that the DSRs given generally line up well with the *Blind* conventional feedback; that is, the DSRs reflect buyer standards similar to those revealed in *Blind*.

To summarize, the *Baseline* treatment qualitatively replicates the pattern of strategic timing, retaliation, and correlation of feedback found on eBay.[21] Moreover, as predicted, the alternative systems successfully mitigate reciprocity (as shown, for instance, by reduced correlations of feedback content) and so allow for a more negative response to lower quality.

4.3. Quality, Prices, and Efficiency

The hypothesis underlying our redesign efforts is that the extent to which feedback is shaped by reciprocity affects economic outcomes. More specifically, we hypothesize that diminishing the role of reciprocity increases quality, prices, and efficiency. Figure 4 shows the evolution of quality and auction prices over time: both quality and prices are higher in both *DSR* and *Blind* than in *Baseline*. Applying a one-tailed Wilcoxon test using independent matching group averages, the increases in average quality and price over all rounds are significant for treatment *DSR* ($p = 0.035$ and 0.025, respectively), but not for *Blind*. The test, however, aggregates over all rounds, and there is a sharp end game effect in all treatments, with both quality and prices falling toward zero, consistent with related studies on reputation building in markets (e.g., Selten and Stöcker 1986). Regressions controlling for round and end game effects yield positive treatment effects regarding quality and prices for both *DSR* and *Blind*, although only *DSR* effects are significant (see Price Model 1 and Quality Model 1 in Table 6).[22]

The choice of bid and quality levels affects efficiency. In the *Baseline* treatment, 47% of the potential value was realized, with losses of 23% and 31% resulting from misallocation and low quality, respectively.[23] Both alternative systems increase efficiency, yet only *DSR* does so significantly; there is a 27% increase in efficiency in *DSR* ($p = 0.027$), compared with *Baseline*, and a 16% increase in *Blind* ($p = 0.320$). Both market sides gain (although not significantly so) in the new system: about 45% (56%) of the efficiency gains end up in the sellers' pockets in *DSR* (*Blind*), and

[20] The same probit, run on all successful auction data (not conditional on the buyer giving feedback), yields similar results, although the coefficient for the *Blind* treatment is somewhat smaller (still positive) but insignificant, most likely because of the drop in feedback frequency we observed earlier for that treatment. The share of positive (negative) buyer-to-seller feedback is 53% (44%) in *CF*, 47% (48%) in *Blind*, and 55% (37%) in *DSR*. Also see the discussion in §4.4 on the informativeness of conventional and detailed seller rating information in *DSR*.

[21] There are two major exceptions. First, there is no endgame effect in the field. Second, we have more negative feedback than eBay. This is desirable because it magnifies the object of our study, feedback retaliation.

[22] Quality Models 1 and 2 in Table 6 reveal another reciprocity effect, resembling what is frequently observed in trust games: sellers respond to higher price offers with better quality (a 1 ECU price increase comes with a 0.2 percentage point increase of quality). Although there is evidence from a controlled field experiment conducted on eBay suggesting that both eBay buyers and sellers may care about reciprocal fairness (Bolton and Ockenfels 2011), we are not aware of any eBay field study investigating whether the final auction price reciprocally affects seller behavior.

[23] A misallocation occurs if the bidder with the highest valuation does not win, so that welfare is reduced by the difference between the highest valuation and the winner's valuation (which we define as the seller's opportunity cost of 100 when there is no winner because of lack of bids). Low quality leads to an efficiency loss because each percent quality the seller does not deliver reduces welfare gains by one percent of the auction winner's valuation, minus one. Also, each feedback reduces welfare by 1 ECU, but this source of efficiency loss is negligible, as no treatment feedback costs exceed 1% of maximal efficiency.

Table 5 Determinants of Problematic Feedback Conditional on Feedback Given, Probit Coefficient Estimates of Marginal Effects (dy/dx)

	Buyer gives problematic feedback			
	Model 1		Model 2	
Dependent variable	Coeff.	(SE)	Coeff.	(SE)
Blind	0.055	(0.030)	0.077**	(0.036)
DSR	−0.029	(0.058)	0.077**	(0.035)
Round	0.001	(0.001)	−0.002**	(0.001)
Price			0.001***	(0.0002)
Quality			−0.009***	(0.001)
S conventional feedback score			−0.006	(0.004)
N	1,725		1,725	
Restricted log-likelihood	−1,183.6		−558.8	

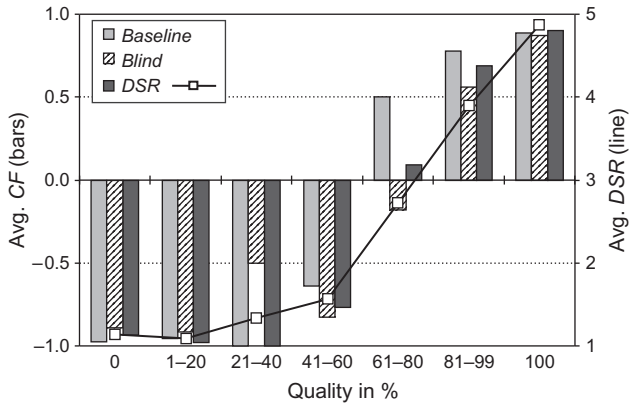
Notes. Robust standard errors clustered on matching group, rounds 1–50. Problematic feedback includes both neutral and negative feedback. *Blind* and *DSR* are treatment dummies. *S conventional feedback score* denotes the feedback score of the seller.

*, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

the rest goes to buyers. So both alternative systems seem to increase price, quality, and efficiency, but only *DSR* improvements are statistically significant. There are changes in the *Blind* treatment, but they are more subtle, as discussed in the next section.

We saw in §4.2 that both proposed systems lead to less reciprocal feedback and in §4.3 that they lead to improved market outcomes. But how does less reciprocity translate into better market performance? The natural hypothesis is that, for a given quality, less reciprocity in feedback giving generates reputation scores that allow better forecasting of sellers' future behavior. In fact, Quality Model 2 in Table 6 shows that sellers' conventional feedback scores in *Blind* have a significantly higher positive correlation with the quality the seller provides at that point than is the case in *Baseline*. The positive correlation between quality and conventional feedback scores increases in *DSR* as well, but not significantly so. Observe, however, that the DSRs are significantly positively correlated with quality and so, in this sense, the *DSR* seller scores, as well as those in *Blind*, exhibit less distortion than those in *Baseline*.

Figure 3 Average Feedback Given After Observing Quality



Note. For this figure, *CF* is coded −1 (negative), 0 (neutral), and 1 (positive). *DSR* is given on a 1–5 integer scale.

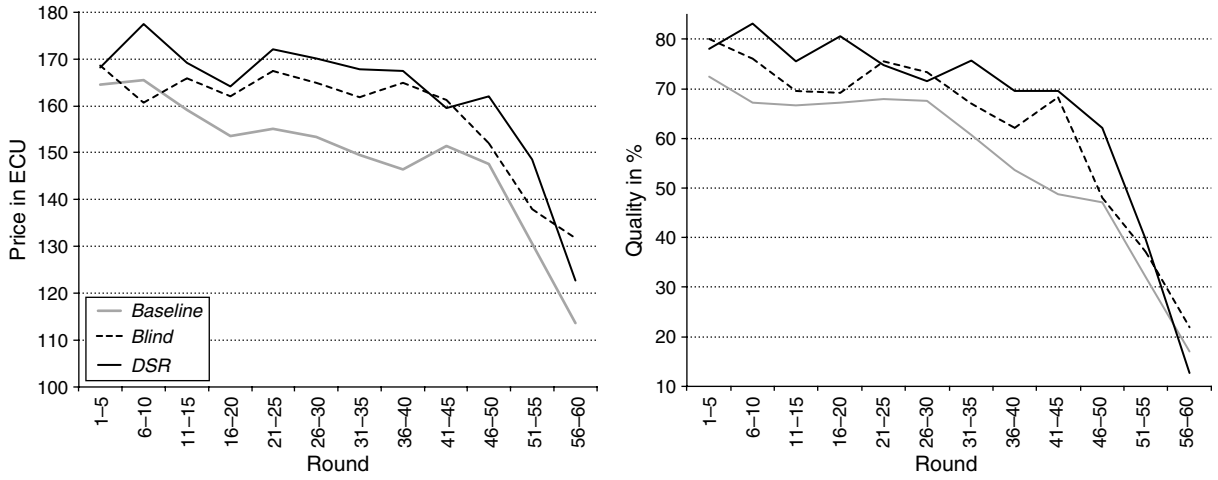
4.4. Relationship Between Feedback Informativeness and Improvement in Market Performance

We expect that introducing one of the alternative systems leads sellers to react to better feedback informativeness by shipping higher quality in *Blind* and *DSR* than in *Baseline*. Returning again to Table 6, Price Model 2 shows that nominally equivalent conventional feedback scores in *Baseline* and *Blind* lead to higher prices in the latter case. In comparing *Baseline* and *DSR*, there is little difference in regard to conventional feedback impact on price; however, DSRs are significantly positively correlated with price, and in this sense sellers with good feedback scores are more highly rewarded in treatment *DSR* than in *Baseline*. More evidence comes from looking directly at the effect a quality decision has on a seller's future average profit. Model 1 in Table 7 shows that the amount of quality a *Baseline* seller chooses in the present round drives up future average profit, but not significantly. In contrast, the amount of quality a *Blind* or *DSR* seller chooses drives up their future expected profit by a higher and significant amount, and by about the same amount for both treatments.[24]

Similarly, a positive feedback score yields stronger incentives to trust in *Blind* and *DSR* than in *Baseline*. For each treatment separately, we ran an ordinary

[24] As a side note, observe that Model 2 in Table 7 shows that knowing a seller's feedback score has greater value for forecasting a seller's future average profit than does knowing the quality decision he makes in the present round, in all three treatments. That is, a summary statistic of a seller's feedback history is a better predictor of his future profitability than observing directly what he did in the present.

Figure 4 Average Auction Prices and Sellers' Quality Choices over Time



least squares regression with buyer payoff as the dependent variable and the conventional feedback score as the sole explanatory variable, and used the estimated model to compute the minimal feedback score such that a risk-neutral buyer is better off trusting the seller (for given prices in each respective treatment). For *Baseline* conventional feedback, the buyer should trust if the feedback score is at least 6. For *Blind* and *DSR* the number falls to 1. So there is good reason to trust a seller with any positive feedback score in *Blind* and *DSR*, but not in the conventional feedback system. The result also suggests that the proposed systems make it easier for new sellers to build up trust and business.

Another way to measure the informativeness of the system is in terms of the amount of confidence we can have that the feedback score reflects the quality that will be received. Specifically, we are interested in the confidence that a high feedback score predicts high quality. We ran a tobit estimation on the *Baseline* data, with quality as the dependent variable and the conventional feedback score as the sole explanatory variable, and used the estimated model to calculate the probability of getting high quality (defined as 90% or greater) conditional on a given feedback score. We then repeated this exercise for *Blind* and *DSR* (again with a conventional feedback score as the sole explanatory variable). The results show that high conventional feedback scores are strongly and about equally more informative in *Blind* and *DSR* than in *Baseline*. For example, a conventional feedback score of 10 implies about a 50% chance of receiving high quality in *Baseline*, but rises to an 80% chance in both *Blind* and *DSR*. Adding the average *DSR* feedback

Table 6 Determinants of Quality and Price, Tobit Coefficient Estimates of Marginal Effects (dy/dx)

	Quality				Price			
	Model 1		Model 2		Model 1		Model 2	
Dependent variable	Coeff.	(SE)	Coeff.	(SE)	Coeff.	(SE)	Coeff.	(SE)
Blind	9.47	(9.182)			9.33	(11.199)		
DSR	20.02***	(6.111)			14.89*	(7.881)		
Round	−0.45***	(0.162)	−1.13***	(0.210)	−0.41**	(0.163)	−1.18***	(0.158)
S FScore			3.85***	(0.977)			5.63***	(0.944)
S FScore ∗ *Blind*			3.38**	(1.696)			3.41***	(0.807)
S FScore ∗ *DSR*			1.05	(1.201)			−0.604	(1.013)
S DSR avg			6.22***	(1.970)			3.75**	(1.847)
Price	0.418***	(0.049)	0.222***	(0.045)				
N	2,283		2,283		2,283		2,283	
Restricted log-likelihood	−8,098.4		−7,933.2		−11,032.8		−11,038.7	

Notes. Robust standard errors clustered on matching group, rounds 1–50. *Blind* and *DSR* are treatment dummies. *S FScore* denotes the (conventional) feedback score of the seller, and *S DSR avg* the seller's average DSR score. Ai and Norton (2003) observe that the tobit interaction effects reported by standard statistical software can be inaccurate because of the nonlinearity of the model. We ran ordinary least squares models with random effects as robustness tests, and the results are largely the same.

*, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Table 7 Determinants of Seller Average Future Profit, Tobit Coefficient Estimates of Marginal Effects (dy/dx)

	Seller average future profit			
	Model 1		Model 2	
Dependent variable	Coeff.	(SE)	Coeff.	(SE)
S FScore			3.04***	(0.489)
S FScore ∗ *Blind*			1.36*	(0.748)
S FScore ∗ *DSR*			−2.30***	(0.763)
S DSR avg			3.92***	(1.083)
Quality ∗ *Baseline*	0.079	(0.083)	−0.019	(0.056)
Quality ∗ *Blind*	0.175**	(0.082)	0.098	(0.062)
Quality ∗ *DSR*	0.179***	(0.0478)	0.034	(0.042)
Nosale	−53.92***	(5.00)	−43.75***	(6.533)
N	2,400		2,400	
Restricted log-likelihood	−11,398.2		−11,180.5	

Notes. Robust standard errors clustered on matching group, rounds 1–50. *Blind* and *DSR* are treatment dummies. *S FScore* denotes the feedback score of the seller, and *S DSR avg* denotes the average DSR score. The period variable is omitted because the associated coefficient is small and insignificant. Ai and Norton (2003) observe that the tobit interaction effects reported by standard statistical software can be inaccurate because of the nonlinearity of the model. We ran ordinary least squares models with random effects as robustness tests, and the results are largely the same.

*, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

score to the *DSR* tobit forecast changes the probability of receiving high quality as *averaged* across all feedback scores little (less than a 1% point increase)—but affects *individual* forecasts by a substantial amount: the average absolute difference between the probability of receiving high quality forecast on the basis of both conventional and *DSR* scores versus that forecast with only conventional feedback is 6.4% points. That is, *DSR* feedback does what we would hope it would do: it separates the performance of traders with the same conventional feedback score.

One interesting question is to what extent the strategic element (blindness) versus the granularity element (five- versus three-point scale) accounts for the increased informativeness of *DSR* feedback relative to baseline conventional feedback. A related question concerns the comparative effectiveness of making *DSR* versus conventional feedback blind. To answer these questions, we reduced the *DSR* feedback given by buyers to a three-point scale: Scores of 4 or 5 were recoded as "positive," 3 as "neutral," and 1 or 2 as "negative," stripping the *DSR* information of its granularity advantage over conventional feedback. We then used the recoded scores to compute total feedback scores in the same way that they are calculated using conventional feedback (see §2.1), producing an encapsulation of *DSR* information that is directly comparable to the blind conventional feedback score, and holding the method of combining past trading information constant. Adding this as a variable to the Quality Model 2 of Table 6, we can distinguish between the strategic effect of *DSR* (as accounted for by the new variable) and the granular effect (now accounted for by the *S DSR avg* variable). We find both *DSR* variable coefficients to be positive and significant at the 1% level. Hence, both strategic and granular factors contribute to the informativeness of *DSR* information. The coefficient estimate for the strategic *DSR* variable is nominally larger than that for the conventional feedback blind variable (*S FScore* ∗ *Blind* in Table 6), although the two effects are not significantly different at any standard test level. Hence, making *DSR* blind is as effective at encouraging informativeness, along a strategic dimension, as is making conventional feedback blind.

Regarding design implications of our lab study, *DSR* yields significant efficiency gains over the *Baseline* treatment and does not decrease feedback frequency. Because this is not the case for *Blind*, at least not significantly, the experiment suggests that *DSR* might be the better option for a system change, if buyer moral hazard is of minor concern. The fact that, in the natural environment, there might also be path dependency and strategic delay issues with a blind system reinforces this conclusion.

4.5. What the Combined Laboratory and Field Data Tell Us

We close this section with some remarks on the complementary relationship between the experiment and the field data examined in §3. From the field evidence, we concluded that the feedback in markets with a blind feedback system exhibits the more disperse, more critical pattern that we desire to achieve in the eBay system. At the same time, how relevant these observations are to the proposed systems being evaluated is questionable, primarily because it is difficult to draw firm casual inferences linking the feedback system to larger market performance or even to the feedback dispersion. The experiment enables us to test both proposed systems on a level playing field. Moreover, we can clearly link the changes in the feedback system to the observed increases in feedback informativeness, in bids, in delivered product quality, and, finally, to overall market efficiency. It is important to note that this clarity is attributable to the tight control the laboratory environment affords, and that this control is achieved by abstracting away from the field environment (see, e.g., Bolton and Ockenfels 2012, Plott 1994, Grether et al. 1981, Kagel and Roth 2000, Chen 2005, Kwasnica et al. 2005, Chen and Sönmez 2006, and Brunner et al. 2010 for similar arguments in the context of auctions, matching, and other markets).

In contrast, if we look solely at the experimental evidence, we see clear evidence of the hypothesized

links between the feedback systems being proposed and improvements in feedback dispersion and market performance. Yet, precisely because the experiment abstracts away from many field complexities, the applicability of the results to the eBay market is questionable. Here the field evidence provides reassurance. At a broad level, we have field examples of feedback systems with performance that is consistent with what we see in the experiment. And at the detailed level, as we previously pointed out, the experiment mirrors many qualitative feedback patterns in the field.

To summarize, this largely consistent picture of how feedback systems affect feedback patterns, the field and lab data improve our confidence that we understand how each proposed system, if implemented, would perform. Of course, not all the gaps have been filled. Three gaps arise from simplifying features in our laboratory context that might turn out to be important in the field. First, the one-sided DSR feedback raises the potential of *buyer* retaliation, which is important if buyer moral hazard turns out to be a significant concern. The experiment abstracts away from buyer moral hazard, on the assumption it will not be significant, and so does not speak to this concern. Second, in the CF system, feedback is posted immediately, so that other buyers can immediately be warned against a seller who becomes untrustworthy (Cabral and Hortaçsu 2010). In contrast, under a blind system, traders can delay the posting of feedback until the feedback deadline is reached and use this time to deceive buyers. Because our laboratory study did not allow this kind of strategy, the blind system's performance might be overestimated. Third, in our experiment all buyers are sometimes sellers and so should care about feedback earned in the buyer role. Although this also holds for a majority of eBay traders (§2.1), those eBay buyers who do not plan to become sellers might be less concerned with negative feedback. As a result, our experiment might overestimate the potential for improvement.

5. A First Look at the Field Implementation of Detailed Seller Ratings

eBay decided to go for a detailed seller rating feedback system under the name "Feedback 2.0" in spring 2007.[25] Under Feedback 2.0, in addition to the conventional feedback, buyers can leave ratings in four dimensions on a five-point scale. These dimensions are as follows: "How accurate was the item description?" "How satisfied were you with the seller's communication?" "How quickly did the seller ship the item?" "How reasonable were the shipping and handling charges?" For each of these ratings, only the number of feedback entries and the average rating are displayed on the seller's feedback page, and only after the seller receives at least 10 ratings.[26] On the feedback submission page, eBay emphasizes that only averages and no individual DSRs can be observed. As a result, DSR is not only blind (in the sense that it cannot be responded to) and one sided (only buyers can give detailed ratings), but also anonymous (sellers cannot identify the DSR provider). In this section we present early evidence on the performance of the new system.

Before we get to the performance of detailed seller ratings (the concern of our main hypotheses), we first look at the performance of conventional feedback (which we did not hypothesize on). Based on our data set 1, Figure 5 shows the share of positive (left y-axis) as well as neutral, negative, and eventually withdrawn feedback (right y-axis) for the last 30 weeks before and the first 10 weeks after introduction of DSR in early March 2007 (vertical dashed line). The shares are quite stable, with the exception of the kink about 10 weeks before the system change, which falls into the preholiday shopping time known for high expectations and time pressure. From the week before to the week after DSR introduction we observe a small drop in positive and an accompanying rise in neutral feedback. This is in line with the experimental results on the DSR system, where we also observe a shift from positive to problematic (in particular, neutral) feedback.[27] However, these changes are small compared to the holiday shock and overall variance and seem not to be persistent, at least for positive feedback. We also do not observe significant changes in CF-giving frequency, timing or correlation between the prechange data set 1 and postchange data set 2 collected in June 2007. We conclude that, overall, there are no or at best small short-term effects in CF because of the introduction of DSR.

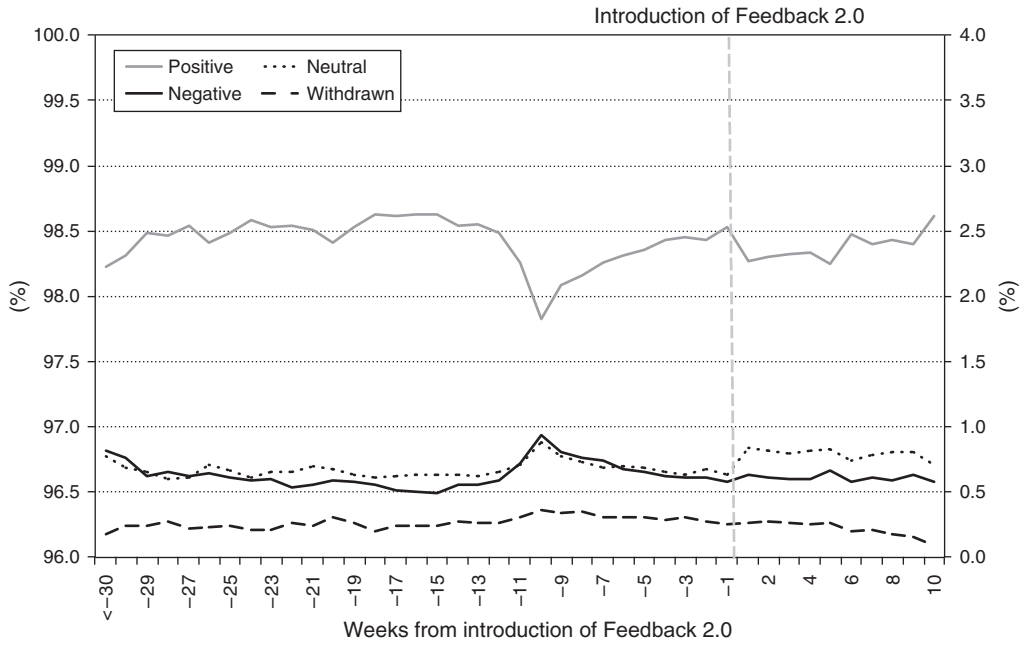
DSRs are given in about 70% of the cases where CF is given, varying somewhat by country and category. Further analysis of conditional DSR frequencies (see

[25] eBay piloted the new design in smaller and medium size eBay markets from early March 2007 (Australia, Belgium, France, India, Ireland, Italy, Poland, and the United Kingdom) and introduced it worldwide in the first week of May 2007.

[26] See Figures 8 and 9 in Online Appendix C for screenshots. Similar to conventional feedback, DSRs are averaged for each buyer before being aggregated. Also, DSRs older than 12 months are ignored, yielding a "rolling" average. There are a number of other small changes implemented jointly with Feedback 2.0. For instance, information about item title and price were added to feedback comments the seller received.

[27] We caution that comparison of lab and field data might be diluted because the field data may reflect transitional dynamics. The experiment did not study the migration from one to another system.

Figure 5 Evolution of Positive, Neutral, Negative, and Withdrawn Feedback Before and After Introduction of Feedback 2.0



Notes. This figure is based on approximately seven and three million individual feedbacks in the 30 weeks before and the first 10 weeks after introduction of Feedback 2.0, respectively, in the pilot countries Australia, Belgium, France, Poland and the United Kingdom. Positive feedback is plotted on the left y-axis, all other feedback on the right y-axis.

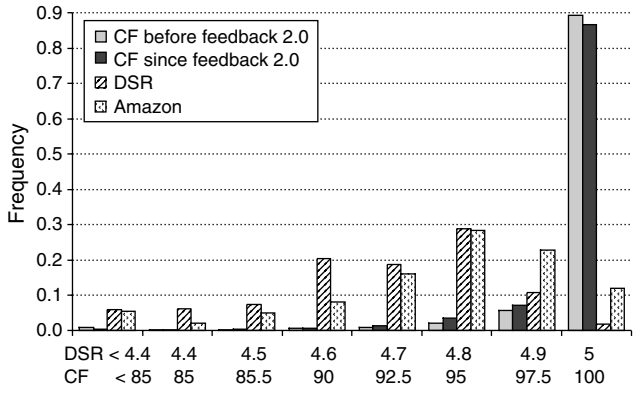
Table 10 in Online Appendix C) shows that DSR feedback is given least often (64%–71%) if the CF feedback is negative and most often (77%–79%) if the CF is neutral.

For the 27,759 eBay members from Australia, Belgium, France, Poland, and the United Kingdom in data set 1, who received at least 10 DSRs between the first week of March and data collection in May 2007 (such that their DSR average was published on their feedback profile), we track CF received as a seller in the same period as well as in the two months before DSR introduction (using individual feedback data from data set 1). From this feedback we calculated the fictitious percentage positives of CF of each individual seller before and after introduction of Feedback 2.0, using only feedback given in the corresponding time windows.

In line with Figure 5, Figure 6 shows that the CF percentage positives scores slightly decreased after introduction of the new system. However, DSRs are more nuanced. For instance, although most sellers have a "perfect" CF reputation of 100%, very few have a "perfect" average DSR of 5. For comparison we also include the distribution of average scores of Amazon.com marketplace sellers in Figure 6 (based on data set 5; see Online Appendix B). The one-sided DSR feedback distribution follows the one-sided Amazon.com feedback distribution fairly closely, although it seems to be somewhat more negative. This supports the idea that DSR is treated as a one-sided system, with little scope for reciprocity. In fact, Figure 7 shows that the difference in rating variability between CF and DSR is partly driven by a strategic response to the differences in the scope for reciprocal behavior.

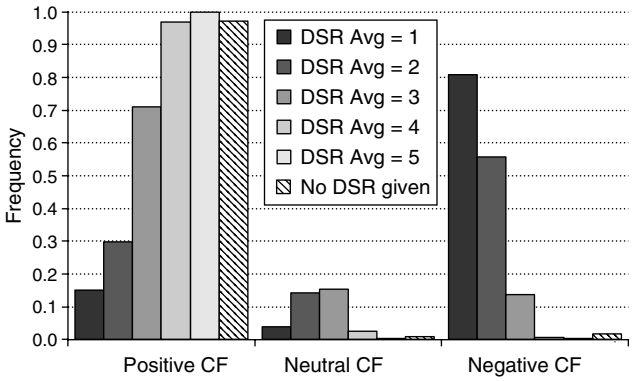
Figure 7 (based on data set 2; see Online Appendix B) shows for each DSR (averaged over the four categories) the distribution of the corresponding CFs.

Figure 6 Distribution of Average CF and DSR Scores in Member Profiles



Notes. DSR and Amazon.com's 1–5 range and CF percent positive's 0–100 range are divided in the same number of categories and are aligned at the x-axis. eBay data are based on the feedback of the same 27,759 members from Australia, Belgium, France, Poland, and the United Kingdom, received as seller in January/February 2007 and March/April/May 2007, respectively. Inclusion criterion was more than 10 DSRs in at least one DSR category. Amazon data are based on 9,741 Amazon marketplace sellers.

Figure 7 Distribution of CF Conditional on Average of Corresponding DSRs



Notes. To calculate the DSR average, we take all available of the up to four DSR ratings per feedback, average, and round to integer. Thus, a DSR average of 1 implies two or three ratings of 1 and at most one rating of 2.

As one might expect, when the DSR is 5, virtually all CF is positive, and when the DSR is 4, almost all CF is positive. However, of those buyers who submit the minimum DSR average of 1 (that means that in the 1–4 DSR ratings, the buyer either gave only 1s or at least two 1s and at most one 2), about 15% submit a positive CF. For DSR averages of 2, this share is 30%. That is, among those who are maximally unsatisfied as measured by DSR, which cannot be reciprocated, a substantial share expresses satisfaction with respect to CF, which can be reciprocated. It seems plausible that at least part of this pattern can be interpreted as hiding bad detailed seller ratings behind positive open conventional feedback.

The initial concern that this kind of strategic hiding behavior might yield inconsistencies between aggregate CFs and DSRs is not borne out, however. The overall share of DSR averages of 1 or 2 is only slightly less than 2%, so that on average, a positive CF comes with a better DSR.

Strategic feedback hiding is only effective when the seller is not able or willing to retaliate against such feedback. However, although DSR makes retaliation more difficult, one might still suspect that by permanently observing the changes of average ratings, a seller could be able to identify the buyer behind a given DSR. This hypothesis is not supported by our data. When the buyer gives an average DSR of 1 but a positive CF, the probability that the seller retaliates upon this with a negative CF is 0.004, compared with a retaliation probability of 0.468 when the CF is negative.[28]

The experiment suggests that most of the endogenous improvement in performance can be expected from pickier buying. In fact, there is evidence indicating that buyers are indeed more distinguishing under DSR. That is, sellers with a relatively good DSR score have a higher probability of selling listed items after introduction of DSRs than the same sellers before introduction of DSRs, and sellers with a relatively low DSR score have a lower probability of selling with DSRs.[29]

6. Conclusions and Challenges for Future Research

This study is a first exploration of the market design issues surrounding the engineering of trust and trustworthiness in the marketplace. It illustrates how gaming in the production of reputation information can significantly hamper the ability of a reputation system to facilitate trust and trade efficiency. Our analysis began with the observation that reciprocity plays a major role in the leaving, timing, and content of feedback. Although retaliatory feedback is in itself a rather small phenomenon, accounting for less than 1.2% of the total mutual feedback data (Figure 1), the *threat* of retaliatory negative feedback distorts feedback in the aggregate. This is because buyers respond strategically to the threat, either by not reporting bad experiences or by waiting for the seller to report first. This in turn reduces the informativeness of feedback, with the end result that a seemingly small phenomenon can substantially hamper trust and market efficiency.

Our study also speaks to the need for new theory. Our method was to observe the phenomenon in the field as best as we could, to design a laboratory study to probe the phenomenon in greater detail and to establish causalities in the laboratory, and then to draw analogies based on the robust findings from field and experimental data. These analogies put the phenomenon in sharper relief and suggest data regularities and questions that any theory of the phenomenon will want to address. A model could usefully describe, for instance, the noise in feedback, such that more compressed feedback makes it harder to correctly predict quality from the reputation score. Moreover, it would also be useful to endogenize the degree of reciprocity in feedback giving in different institutional environments. This may involve utilizing models of reciprocity, social comparison, and group identity (for related observations, see Chen et al. 2010, Chen and Li 2009). Combining theory and empirical

[28] In support of this observation, straightforward regression analyses show a very high correlation between seller's and buyer's CF, but when controlling for the buyer's CF, correlations with DSRs are very low, or even negative.

[29] The effects are statistically highly significant. The data necessary to document this cannot be presented here because eBay asked that it be kept confidential (the only data eBay did not allow us to use in this paper).

studies will further improve our understanding of the role of behavior and design in reputation building.

Our study also has implications for managing the redesign of market trust systems. First, a major challenge in solving marketplace trust problems has to do with possible adverse side effects or disruptions in path dependencies in migrating to a new system. For example, a redesign of a trust system needs to respect the fact that reciprocity has positive as well as negative consequences for the feedback system. The giving of feedback is largely a public good, and our data suggest that reciprocity is important for recording mutually satisfactory trades. It is therefore desirable that, in mitigating retaliatory feedback, we strive for a targeted approach rather than one that attempts to remove all forms of reciprocity. Also, by nature, reputation mechanisms are embedded in repeated games, connecting past with future behavior. It was important to the present redesign to maintain certain aspects of the old system, such as the three-point (conventional) scoring, so that the information collected prior to the change in the system would still be useful in evaluating traders after the changeover, without causing undue confusion.

Second, our laboratory study shows that reciprocal feedback behavior can be channeled, and in a targeted way. The way feedback information is navigated through the system affects whether and how reciprocity influences the candor of feedback. The data show that, compared to a simple open system, both blindness in conventional feedback giving and one-sidedness in a detailed seller rating system increase the information contained in the feedback presented to buyers. As a result, the redesigns likely yield more trust and efficiency in the market, at least in the short-run period that we studied. Additional studies, particularly of longer-term effects, should yield further insights.

A third implication has to do with the strength of approaching the problem using complementary methods of analysis. It is the combination and complementary nature of the lab and field data that allows us to be confident in our judgment of likely consequences of institutional changes. With only field data, it would be difficult to establish the influence of institutional differences, because both cross- and within platform comparisons involve confounding environment factors. At the same time, laboratory experiments alone do not capture various complexities of the corresponding field environments.[30] They do, however, demonstrate (beyond their benefits as a test bed for competing designs) that the interaction of institutions and reciprocal behaviors is sufficient to produce the robust empirical patterns observed in the various data sets. And when taken together, our laboratory and field investigations of different feedback systems provide a surprisingly coherent picture of how institutional change affects social and strategic feedback giving.

eBay introduced detailed seller ratings in March and May 2007. Relative to the conventional feedback on eBay, this feedback is more detailed, one sided, and anonymous. The change did not much affect CF giving, but many traders use the new system to avoid retaliation. This contributes to more reputation dispersion, which in turn leads to improved informativeness.

Naturally, market platforms like eBay continuously monitor and improve trust and trustworthiness on their platforms. Motivated by the positive effects of detailed seller ratings, eBay moved ahead and introduced further changes in spring 2008. The most important feature of this more recent change is that sellers are not allowed to submit negative or neutral feedback anymore, only positive. Basically, this is a move to a one-sided feedback system, as found on many business-to-consumer platforms, but it still allows for positive reciprocity. Further research will be devoted to how this new change affects the content, timing, and informativeness of feedback. For example, one might expect that, contrary to their behavior in the previous design, more sellers will move first in feedback giving to trigger positive reciprocity.

There are other important challenges in designing feedback systems that are not addressed here. For example, reputation profiles may be tradable, such that new sellers may buy their reputation from the market (see Brown and Morgan 2006). Or traders might change their online identity or maintain multiple profiles. These factors too may undermine

[30] Laboratory engineering studies can be done on different levels of attention to detail. Our study was not designed to maximally emulate the eBay environment. For instance, eBay's DSRs are less prominently displayed on eBay (the DSRs are one click away) than in the laboratory experiment. To the extent the presentation is fixed and affects long-run behavior and performance, there is a need to qualify our experimental results for the eBay context. If, for instance, traders do not use DSRs on eBay because they do not see them, our results would overestimate the benefits of DSRs. However, basing recommendations on experience in complicated ("close to real-world") environments comes at a cost. The presentation of the feedback system, for instance, can be quickly changed: if DSRs work better than expected, eBay can decide to place them more prominently. This is why this detail should not affect the basic decision of whether to change the system. Similarly, we decided to choose a sealed-bid format that abstracts away from eBay's bidding dynamics, which we did not consider to be of much relevance with respect to feedback dynamics. The experiment design sought a level of detail that does not sacrifice the environmental features relevant to the purpose of study, but at the same time is general enough to generate robust insights into the effectiveness of different feedback systems to diminish retaliation and increase feedback informativeness.

the informativeness of a feedback system. The engineering approach might usefully be applied to sort potential solutions here as well.

Acknowledgments

The authors thank Brian Burke, Debbie Hofmeyr, Leland Peterson, Joe Fuller, and the rest of eBay's Trust and Safety Team for their cooperation in this project. The authors provided eBay with advice on improving their feedback system under an agreement that permitted eBay site data to be used for academic publication. The authors also thank two anonymous referees and the anonymous associate editor for very helpful advice, as well as Al Roth and Ido Erev for helpful comments. The authors are indebted to Felix Lamouroux, Karin Ruetz, and Dietmar Ilsen for excellent research assistance and help with the collection of data. Financial support from the German Science Foundation through the Leibniz program and the research group Design & Behavior and from the U.S. National Science Foundation [Grant 0351408] is gratefully acknowledged.

References

Ai C, Norton EC (2003) Interaction terms in logit and probit models. *Econom. Lett.* 80(1):123–129.

Akerlof G (1970) The market for lemons: Quality uncertainty and the market mechanism. *Quart. J. Econom.* 84(3):488–500.

Ariely D, Ockenfels A, Roth AE (2005) An experimental analysis of ending rules in Internet auctions. *RAND J. Econom.* 36(4):790–809.

Ba S, Pavlou P (2002) Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quart.* 26(3):243–268.

Bajari P, Hortaçsu A (2003) The winner's curse, reserve prices and endogenous entry: Empirical insights from eBay auctions. *RAND J. Econom.* 34(2):329–355.

Bajari P, Hortaçsu A (2004) Economic insights from Internet auctions. *J. Econom. Literature* 42(2):457–486.

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

Bolton G, Ockenfels A (2009) The limits of trust in economic transactions: Investigations of perfect reputation systems. Cook KS, Snijders C, Buskens V, Cheshire C, eds. *eTrust: Forming Relationships in the Online World* (Russell Sage, New York), 15–36.

Bolton G, Ockenfels A (2011) Does laboratory trading mirror behavior in real world markets? Fair bargaining and competitive bidding on eBay. Working paper, University of Cologne, Cologne, Germany.

Bolton G, Ockenfels A (2012) Behavioral economic engineering. *J. Econom. Psych.* 33(3):665–676.

Bolton G, Katok E, Ockenfels A (2004) How effective are online reputation mechanisms? An experimental study. *Management Sci.* 50(11):1587–1602.

Bolton G, Katok E, Ockenfels A (2005) Cooperation among strangers with limited information about reputation. *J. Public Econom.* 89(8):1457–1468.

Bolton GE, Ockenfels A, Ebeling F (2011) Information value and externalities in reputation building. *Internat. J. Indust. Organ.* 29(1):23–33.

Brandts J, Figueras N (2003). An exploration of reputation formation in experimental games. *J. Econom. Behav. Organ.* 50(1): 89–115.

Bretz RD, Milkovich GT, Read W (1992) The current state of performance appraisal research and practice: Concerns, directions, and implications. *J. Management* 18(2):321–352.

Brown J, Morgan J (2006) Reputation in online markets: The market for trust. *California Management Rev.* 49(1):61–81.

Brunner C, Goeree JK, Holt CA, Ledyard JO (2010) An experimental test of combinatorial FCC spectrum auctions. *Amer. Econom. J.: Microeconomics* 2(1):39–57.

Cabral L, Hortaçsu A (2010) The dynamics of seller reputation: Evidence from eBay. *J. Indust. Econom.* 58(1):54–78.

Camerer CF (2003) *Behavioral Game Theory* (Princeton University Press, Princeton, NJ).

Camerer CF, Weigelt K (1988) Experimental tests of a sequential equilibrium reputation model. *Econometrica* 56(1):1–36.

Chen K (2005) An economics wind tunnel: The science of business engineering. Morgan J, ed. *Experimental and Behavioral Economics—Advances in Applied Microeconomics*, Vol. 13 (Elsevier Press, New York).

Chen Y, Li SX (2009) Group identity and social preferences. *Amer. Econom. Rev.* 99(1):431–457.

Chen Y, Sönmez T (2006) School choice: An experimental study. *J. Econom. Theory* 127(1):202–231.

Chen Y, Harper M, Konstan J, Li SX (2010) Social comparisons and contributions to online communities: A field experiment on MovieLens. *Amer. Econom. Rev.* 100(4):1358–1398.

Chwelos P, Dhar T (2007) Differences in "truthiness" across online reputation mechanisms. Working paper, Sauder School of Business, University of British Columbia, Vancouver.

Cooper D, Kagel J (2012) Other-regarding preferences. Kagel J, Roth A, eds. *The Handbook of Experimental Economics*, Vol. 2 (Princeton University Press, Princeton, NJ). Forthcoming.

Dellarocas C (2004) Building trust on-line: The design of robust reputation mechanisms for online trading communities. Doukidis G, Mylonopoulos N, Pouloudi N, eds. *Social and Economic Transformation in the Digital Era* (Idea Group Publishing, Hershey, PA), 95–113.

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

Dewan S, Hsu V (2001) Trust in electronic markets: Price discovery in generalist versus specialty online auctions, Working paper, University of Washington, Seattle.

Dulleck U, Kerschbamer R, Sutter M (2011) The economics of credence goods: On the role of liability, verifiability, reputation and competition. *Amer. Econom. Rev.* 101(2):526–555.

Eaton DH (2007) The impact of reputation timing and source on auction outcomes. *B.E. J. Econom. Anal. Policy* 7(1), Article 33.

Ederington LH, Dewally M (2006) Reputation, certification, warranties, and information as remedies for seller-buyer information asymmetries: Lessons from the online comic book market. *J. Bus.* 79(2):693–729.

Fehr E, Gächter S (2000) Fairness and retaliation: The economics of reciprocity. *J. Econom. Perspectives* 14(3):159–181.

Greif A (1989) Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *J. Econom. History* 49(4):857–882.

Greiner B (2004) An online recruitment system for economic experiments. Kremer K, Macho V, eds. *Forschung und wissenschaftliches Rechnen 2003*, GWDG Bericht 63 (Ges. für Wiss, Datenverarbeitung, Göttingen), 79–93.

Greiner B, Ockenfels A, Sadrieh A (2012) Internet auctions. Peitz M, Waldfogel J, eds. *Oxford Handbook of the Digital Economy* (Oxford University Press, New York), 306–342.

Grether DM, Isaac RM, Plott CR (1981) The allocation of landing rights by unanimity among competitors. *Amer. Econom. Rev.* 71(2):166–171.

Grosskopf B, Sarin R (2010) Is reputation good or bad? An experiment. *Amer. Econom. Rev.* 100(5):2187–2204.

Güth W, Mengel F, Ockenfels A (2007) An evolutionary analysis of buyer insurance and seller reputation in online markets. *Theory Decision* 63(3):265–282.

Herrmann B, Thöni C, Gächter S (2008) Antisocial punishment across societies. *Science* 319(5868):1362–1367.

Houser D, Wooders J (2005) Reputation in auctions: Theory and evidence from eBay. *J. Econom. Management Strategy* 15(2):353–369.

Jian L, MacKie-Mason J, Resnick P (2010) I scratched yours: The prevalence of reciprocation in feedback provision on eBay. *B.E. J. Econom. Anal. Policy* 10(1), Article 92.

Jin GZ, Kato A (2006) Price, quality and reputation: Evidence from an online field experiment. *RAND J. Econom.* 37(4):983–1005.

Kagel JH, Roth AE (2000) The dynamics of reorganization in matching markets: A laboratory experiment motivated by a natural experiment. *Quart. J. Econom.* 115(1):201–235.

Kalyanam K, McIntyre S (2001) Returns to reputation in online auction markets. Retail Workbench Working Paper W-RW01-02, Santa Clara University, Santa Clara, CA.

Klein TJ, Lambertz C, Spagnolo G, Stahl KO (2007) Last minute feedback. Working paper, University of Mannheim, Mannheim, Germany.

Kreps DM, Wilson R (1982) Reputation and imperfect information. *J. Econom. Theory* 27(2):253–279.

Kwasnica AM, Ledyard JO, Porter D, DeMartini C (2005) A new and improved design for multiobject iterative auctions. *Management Sci.* 51(3):419–434.

Lewis G (2011) Asymmetric information, adverse selection and online disclosure: The case of eBay motors. *Amer. Econom. Rev.* 101(4):1535–1546.

Livingston JA (2005) How valuable is a good reputation? A sample selection model of Internet auctions. *Rev. Econom. Statist.* 87(3):453–465.

Livingston JA, Evans WN (2004) Do bidders in Internet auctions trust sellers? A structural model of bidder behavior on eBay. Working paper, Bentley University, Waltham, MA.

Lucking-Reiley D, Bryan D, Prasad N, Reeves D (2007) Pennies from eBay: The determinants of price in online auctions. *J. Indust. Econom.* 55(2):223–233.

McDonald CG, Slawson Jr, VC (2002) Reputation in an Internet auction market. *Econom. Inquiry* 40(3):633–650.

Melnik MI, Alm J (2002) Does a seller's reputation matter? Evidence from eBay auctions. *J. Indust. Econom.* 50(3):337–349.

Milgrom P (2004) *Putting Auction Theory to Work* (Cambridge University Press, Cambridge, UK).

Milgrom P, North D, Weingast B (1990) The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Econom. Politics* 2(1):1–23.

Miller N, Resnick P, Zeckhauser R (2005) Eliciting informative feedback: The peer prediction method. *Management Sci.* 51(9): 1359–1373.

Muniz J, Garcia-Cueto E, Lozano LM (2005) Item formation and the psychometric properties of the eysenck personality questionnaire. *Personality Individual Differences* 38(1):61–69.

Neral J, Ochs J (1992) The sequential equilibrium theory of reputation building: A further test. *Econometrica* 60(5):1151–1169.

Niederle M, Roth AE (2005) The gastroenterology fellowship market: Should there be a match? *Amer. Econom. Rev. Papers Proc.* 95(2):372–375.

Nunnally JC (1978) *Psychometric Theory*, 2nd ed. (McGraw-Hill, New York).

Ockenfels A (2003) Reputationsmechanismen auf Internetmarktplattformen: Theorie und empirie. *Zeitschrift für Betriebswirtschaft* 73(3):295–315.

Ockenfels A, Resnick P (2012) Negotiating reputations. Bolton G, Croson R, eds. *The Oxford Handbook of Conflict Resolution* (Oxford University Press, Oxford, UK), 223–237.

Ockenfels A, Roth AE (2006) Late and multiple bidding in second price Internet auctions: Theory and evidence concerning different rules for ending an auction. *Games Econom. Behav.* 55(2):297–320.

Oppenheim AN (2000) *Questionnaire Design, Interviewing and Attitude Measurement* (Continuum, London).

Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action* (Cambridge University Press, Cambridge, UK).

Özer Ö, Zheng Y, Ren Y (2012) Trust, trustworthiness, and information sharing in supply chains bridging China and the U.S. Working paper, University of Texas at Dallas, Richardson.

Plott CR (1994) Market architectures, institutional landscapes and testbed experiments. *Econom. Theory* 4(1):3–10.

Prendergast C (1999) The provision of incentives in firms. *J. Econom. Literature* 37(1):7–63.

Prendergast C, Topel RH (1993) Discretion and bias in performance evaluation. *Eur. Econom. Rev.* 37(2–3):355–365.

Reichling F (2004) Effects of reputation mechanisms on fraud prevention in eBay auctions. Working paper, Stanford University, Stanford, CA.

Resnick P, Zeckhauser R (2002) Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. Baye MR, ed. *The Economics of the Internet and E-Commerce*, Advances in Applied Microeconomics, Vol. 11 (JAI Press, Amsterdam), 127–157.

Resnick P, Zeckhauser R, Friedman E, Kuwabara K (2000) Reputation systems. *Comm. ACM* 43(12):45–48.

Resnick P, Zeckhauser R, Swanson J, Lockwood L (2006) The value of reputation on eBay: A controlled experiment. *Experiment. Econom.* 9(2):79–101.

Roth AE (2002) The economist as engineer: Game theory, experimentation, and computation as tools for design economics. Fisher-Schultz lecture. *Econometrica* 70(4):1341–1378.

Roth AE (2008) What have we learned from market design? *Econom. J.* 118(527):285–310.

Roth AE, Ockenfels A (2002) Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the Internet. *Amer. Econom. Rev.* 92(4):1093–1103.

Selten R, Stöcker R (1986) End behavior in sequences of finite prisoner's dilemma supergames. *J. Econom. Behav. Organ.* 7(1): 47–70.

Sutter M, Haigner S, Kocher M (2010) Choosing the stick or the carrot?—Endogenous institutional choice in social dilemma situations. *Rev. Econom. Stud.* 77(4):1540–1566.

Wilson R (1985) Reputations in games and markets. Roth AE, ed. *Game-Theoretic Models of Bargaining* (Cambridge University Press, Cambridge, UK), 27–62.