No. 33

UNIVERSITY OF COLOGNE

WORKING PAPER SERIES IN ECONOMICS

THE LIMITS OF TRUST IN ECONOMIC TRANSACTIONS - INVESTIGATIONS OF PERFECT REPUTATION SYSTEMS

GARY E BOLTON
AXEL OCKENFELS

Department of Economics
University of Cologne
Albertus-Magnus-Platz
D-50923 Köln
Germany

http://www.wiso.uni-koeln.de

The Limits of Trust in Economic Transactions

Investigations of Perfect Reputation Systems

Gary E Bolton*
Penn State University, USA

Axel Ockenfels
University of Cologne, Germany

13 October 2006

* Gary Bolton, Smeal College of Business, Penn State University, University Park, PA 16802 USA, gbolton@psu.edu (http://www.personal.psu.edu/faculty/g/e/geb3/). Axel Ockenfels, University of Cologne, Department of Economics, Albertus Magnus Platz, D-50923 Cologne, Germany, ockenfels@uni-koeln.de (http://ockenfels.uni-koeln.de/ao). Bolton gratefully acknowledges the support of the National Science Foundation. Ockenfels gratefully acknowledges the support of the Deutsche Forschungsgemeinschaft. We are advising firms, including eBay, on reputation mechanism design and other market design issues; the views expressed are our own.

I. Introduction: Why study perfect reputation systems?

As the Internet economy has grown, so too has the need for trust. A degree of trust is a critical ingredient in virtually all economic relationships, Internet or otherwise. Every day we choose to trust plumbers, doctors, employers, employees, teachers, airlines, and others. The need for trust arises from the fact that we cannot contract on every move other's make. And what we can contract on is often prohibitively costly to enforce. The anonymity of geographically dispersed Internet traders increases contracting difficulties: You may not be able to identify your eBay seller or verify the quality of the object being sold, let alone get your money back.

The economic foundation of trust relationships is the reciprocity principle of tit-for-tat. Broadly speaking there are two forms. *Direct* reciprocity applies to repeated relationships: 'I will trust you tomorrow if you are trustworthy with me today.' *Indirect* reciprocal systems enforce trust when the relationship is one-shot by a more circuitous tit-for-tat: 'I will trust you tomorrow if you are trustworthy with a third party today.' Internet markets tend to be anonymous places and feature a lot of one-time transactions. A study by Resnick and Zeckhauser (2002), for example, finds that a large majority of eBay trading encounters are one-shot. As a result, Internet markets tend to lean heavily on indirect reciprocity systems to enforce trustworthiness.

Take eBay's famous feedback forum, a kind of 'institutionalized gossip'.[1] On eBay, after each encounter, buyers and sellers can evaluate each other by giving the trading partner either a positive (+1), neutral (0) or negative (-1) feedback (and maybe an additional verbal commentary). These feedbacks are publicly available and easy to access, so that each buyer can look at a seller's feedback history before he engages in bidding. In this way the incentives for moral hazard are weakened by the feedback system: If eBay traders punish sellers with negative feedback information by refusing to buy from them or reducing the price they are willing to pay, then the threat of leaving negative feedback should discipline the seller.

In this paper, we discuss our investigations of 'perfect' reputation systems for indirect reciprocity. By 'perfect' we mean that the information about traders' past behavior circulating

[1] While online auction transactions appear to be particularly vulnerable to fraud, the problems we report here exist in basically *all* reputation-based interaction. EBay is a convenient example because it allows researchers to quantify some of the benefits and problems. We also note that eBay's feedback forum is only one part of a mix of (imperfect) policies and rules that interact to promote trade efficiency. Only very few papers address this interaction. One exception is Güth et al. (2006), who investigate the joint effectiveness of buyer insurance (which is part of eBay's so-called Purchase Protection Program) and eBay's feedback forum.

through the market is comprehensive and reliable. The effectiveness and robustness of reputation systems critically depends on the quality of the reputation information available. In this regard, our investigations give these systems there best chance. However, real world reputation systems are necessarily imperfect, and this raises the question of why we study perfect systems. Towards answering this question it is good to have an understanding of the nature of the imperfections that challenge real world systems.

One challenge for feedback systems such as eBay's is that feedback information must come from voluntary self-reporting of one's own experiences with trading partners. But feedback is a public good; the costs of providing feedback are paid by the provider but the benefit only goes to *other* traders.[2] Furthermore, no trader can be excluded from using the information. As a result, economic theory suggests that feedback information will be underprovided. In fact, only about 50 percent of the transactions on eBay receive feedback (Resnick and Zeckhauser 2002).[3]

A second challenge is that feedback needs to be reliable to effectively deter fraudulent behavior. There are a variety of incentives to manipulate feedback, for example, to give good feedback to friends and bad feedback to competitors. Another major challenge is that negative feedback is often retaliated by negative feedback, creating incentives not to give negative feedback. It appears suspicious that less than half percent of the eBay feedbacks are negative (as observed by Resnick and Zeckhauser 2002, among others). Further evidence for the limited reliability of eBay's feedback information comes from the observation that negative feedbacks are given late, in the 'last minute'. On the other hand, positive feedback tends to be given earlier in order to trigger a reciprocal response (e.g., Klein et al. 2005). As a consequence, the information value of feedback, if given at all, is likely to be something less than perfect.[4]

[2] The cost of generating feedback includes the risk of trusting sellers, something we discuss below.

[3] One of the main motives for giving feedback appears to be reciprocity (Dellarocas et al. 2003). That is, a trader's propensity to leave feedback is driven by the expectation that a positive feedback is reciprocated by the trading partner. This observation is remarkably in line with the literature in experimental economics on voluntary provision of public goods (see, e.g., Ledyard 1995, Ockenfels and Weimann 1999, and the references cited therein).

[4] Another potential source of noisy feedback information is fraudulent identity change. The costs of changing an online trader identity is often close to zero, implying that fraudulent sellers can exploit their buyers and then reappear with a clean record. Ockenfels (2003) shows, however, that if buyers are only willing to buy from a 'newbie' (a seller with no record) if the object is offered at a lower price (compared to the price offered by a seller with a positive reputation record), then trust and trustworthy behavior can be sustained; see Friedman and Resnick (2001) for a similar argument.

So given the real world imperfections, why do we study perfect reputation systems in the laboratory? The first answer to this question is that while most of the empirical literature observes that traders do *respond* to reputation information, this observation does not *measure* the virtues of reputation information. Evidence from eBay, for instance, shows that an eBay-seller's feedback profile may affect prices and the probability of sale (see Dellarocas 2003 and forthcoming, Resnick et al. 2006). The empirical results are mostly consistent with the theoretical expectation of buyers paying more to sellers with better reputations. It has also been observed that the impact of reputation ratings on buyers' behavior tends to be stronger for riskier transactions and more expensive objects. This would seem to indicate that the reputation systems have at least some merit. But precisely how much is gained from these systems in terms of overall cooperation levels and efficiency gains remains unclear. Part of the reason is that most of the economic theory on how these systems perform has been worked out assuming that the feedback systems are perfect, in the institutional sense that they provide comprehensive and reliable information; it is largely unclear to what extent these results are robust to imperfect systems. Another part of the reason is that field studies have difficulties separating imperfect institutions and boundedly rational behavior. It may be that flawed systems work well because real world traders are unable to fully exploit the flaws as theories assuming full rationality would suggest they do. Current studies are discovering changes in rules, procedures and information aggregation that may well help to generate more reliable information. Retaliatory feedback might be eliminated by not letting sellers evaluate buyers as suggested by, e.g., Güth et al. (2006), or by having a 'blind period' in which trading partners can 'simultaneously' leave feedback on each other as suggested by, e.g., Klein et al. (2005). Clever incentive schemes, either based on economics (Miller et al. 2005) or social psychology (Rashid et al. 2006), may overcome the public goods problem and promote full provision of all relevant feedback information. And modern authentication technologies or 'entry fees' may eliminate manipulative changes of online identities (see Friedman and Resnick 2001). But maybe the binding limitation for the effectiveness of reputation systems is not so much the institutional issues but rather the behavioral limitations. By studying perfect reputation systems, we identify the maximal achievable benefit by market design improvements – absent from all kinds of institutional noise and incentive problems inherent to all real world reputation systems – and, by the same token, we reveal how behavioral aspects may limit or assist the reputation system performance. In this

sense, by studying perfect systems, we get a clearer measure of the interplay of institutional and behavioral aspects on the effectiveness of reputation systems.

We study the scope and limitations of perfect reputation systems in thought experiments (using economic theory) and in laboratory experiments (exposing 'real' people to perfect systems). What we find, as we will illustrate below, is that economic theory tends to underestimate traders' intrinsic willingness to behave reciprocally, but at the same time tends to overestimate the effectiveness of extrinsic motivations through reputation institutions. One implication of our work is that understanding how social behavior can be sustained with the help of reputation mechanisms will require a coming to grips with the interplay between the institutional environment and boundedly rational behavior.

II. Intrinsic motivation: What can be achieved without a reputation system?

Standard economic theory, based on a narrow definition of self interest, implies that without external control and incentives, there is hardly any hope that trust and trustworthiness can emerge, but also that a perfect reputation system tends to create sufficient incentives to solve the problems. However, our work suggests that economic theory is misleading on both sides: There can be trust without external enforcement, and there can be cooperation failure even with perfect reputation systems. Thus, when we attempt to measure the impact of the introduction of a perfect reputation system in a community of strangers, we need to carefully measure both, how well the community does absent *any* external cooperation enforcement, and how well it does with a *perfect* enforcement system. While in 'reality' neither environment exists, we can create these situations in the laboratory. For instance, we can create situations, which are anonymous and truly one-shot for our subjects in the sense that none of the encounters is linked by flows of reputation information.

To make things simple and to abstract away various complicating factors in our thought and laboratory experiments, we focus on a simple buyer-seller game that features a trust problem typical of the types of problems that reputation systems are commonly used to mitigate. Figure 1 illustrates the moves in the buyer-seller encounter. Both the seller and the buyer are endowed with 35, which is the payoff when no trade takes place. The seller offers an item for sale at a price of 35 which has a value of 50 to the buyer. The seller's cost of providing the item is 20. If the buyer chooses to buy the item, he sends his endowment of 35 to the seller, who then has to

decide whether to ship the item, or whether to keep both the money and the item. If the seller does not ship, he receives the price plus his endowment of 35 for a total of 70. If he ships, he receives the price minus the costs plus his endowment for a total of 50. If the buyer chooses not to buy the item, no trade occurs.

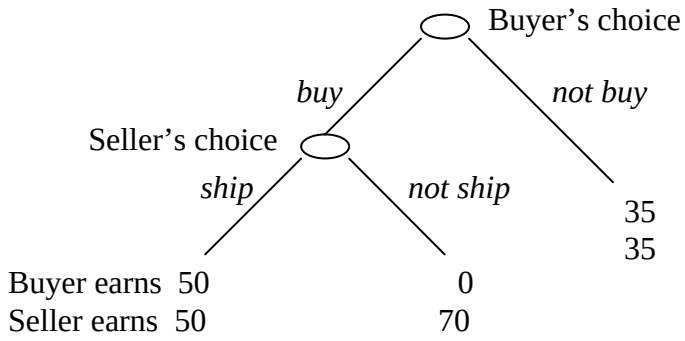

Figure 1. The buyer-seller encounter

At the heart of the game is a moral hazard problem that must be overcome if trades are to be successfully executed. With no common history or common future among traders that could give them the opportunity to reward or punish each other, and with no other kind of external (say, legal) enforcement, the seller can profit from not sending the item or sending poorer quality than promised. That is, the seller's pecuniary motive in the game of Figure 1 dictates to keep the money along with his endowment. In this case, the buyer would lose his endowment and end up with nothing. Anticipating this moral hazard, buyers may not be willing to buy, and thus no transaction will take place. As a consequence, trading that would make everybody better off would not take place. This is the essential trust dilemma that economic and social – whether they be online or offline – interactions need navigate.[5]

[5] We assume that the seller fixes the price. For example, Amazon.com permits sellers of used books and cds to make offering on its site (along with its own new goods offerings). A used goods seller posts an offer on the market platform that includes a description of the item and its condition, and a price at which he is willing to sell. A willing buyers sends the money to Amazon. Upon receiving the money, the seller is supposed to ship the item to the buyer. (In addition, the moral hazards surrounding shipping and accurate representation of good quality are controlled by a feedback system not unlike the one we will introduce to our game.) However, all arguments in this paper hold equally if the price is endogenously determined, such as in eBay's auctions (in this case the auction winner is the buyer).

Economic theory presumes that all rational sellers will fall to moral hazard under the laid out circumstances, and consequently all trustworthiness, and therefore trust, will vanish. However, the standard models assume that people are guided solely by pecuniary concerns. In reality people care about other things as well. In fact, in trust games and related anonymous one-shot games (like the prisoner's dilemma game and the ultimatum game), psychologists, sociologists, experimental economists and others have identified several non-pecuniary motives that are important drivers of behavior in these situations. Most prominently in the recent economics literature are concerns for fairness (Fehr and Schmidt, 1999, Bolton and Ockenfels, 2000) and reciprocity (Rabin, 1993, Dufwenberg and Kirchsteiger, 2004). These social preference models assume that traders care about their own monetary payoff but that some traders may additionally be concerned with the social impact of their behavior. Reciprocity models conjecture that people tend to be kind in response to kindness and unkind in response to unkindness, while fairness models posit that some individuals may have a preference for equitably sharing the efficiency gains from trade.[6]

We studied the game in Figure 1 in a classroom experiment (Bolton, Katok and Ockenfels 2004b), and we found that 37 percent of the 30 sellers were willing to ship in anonymous one-shot encounters. 27 percent of the buyers were willing to buy. So contrary to the predictions of standard theory, there is a non-trivial amount of trust and trustworthiness in anonymous one-shot encounters.[7] At the same time, however, there is substantial room for improvement. On average, only about 10 percent of all encounters (0.27*0.37) end up in successful and efficient trade. Furthermore, this figure probably overestimates the power of intrinsic motivations to behave reciprocally in a dynamic setting: In expected monetary terms, the probability of a trustworthy seller needs to be at least 70% in order to make buying in the trust game profitable. In our one-shot game, the probability was well below this threshold. A natural hypothesis, then, is that if trust rests solely on behavioral propensities, trust will diminish over time.

[6] To be more specific, in our trust game, reciprocity models suggest that a seller ships because the buyer was so *kind* to buy, whereas fairness models suggest that he ships because otherwise the payoff *distribution* would be unfair. See also the discussion of motives like efficiency-concerns and procedural fairness in related games in Bolton and Ockenfels (forthcoming).

This hypothesis has been tested by Bolton, Katok and Ockenfels (2004a). In our laboratory experiment, the market transactions take place over a series of 30 rounds. At the beginning of each round, a potential buyer is matched with a potential seller and they then play the trust game in Figure 1. Each game is played with a different transaction partner and no information about trade outcomes leaks from one encounter to another one, so we call this experimental treatment the *Strangers market*. All interaction was computer-mediated and anonymous; subjects sat in cubicles in front of computers not knowing the true identity of their trading partners, capturing an important aspect of online trading. The rules, and that all rounds would be paid, were common knowledge. Observe that, absent reputation information, this market is essentially a sequence of one-shot games. Thus, because there is not sufficient intrinsic trustworthiness to make trust profitable in the non-repeated one-shot game, we hypothesize that buyers quickly learn that cooperation does not pay out and that, subsequently, trading activities will collapse.

Figure 2 shows the average buying and shipping (conditioned on buying) behavior across rounds.

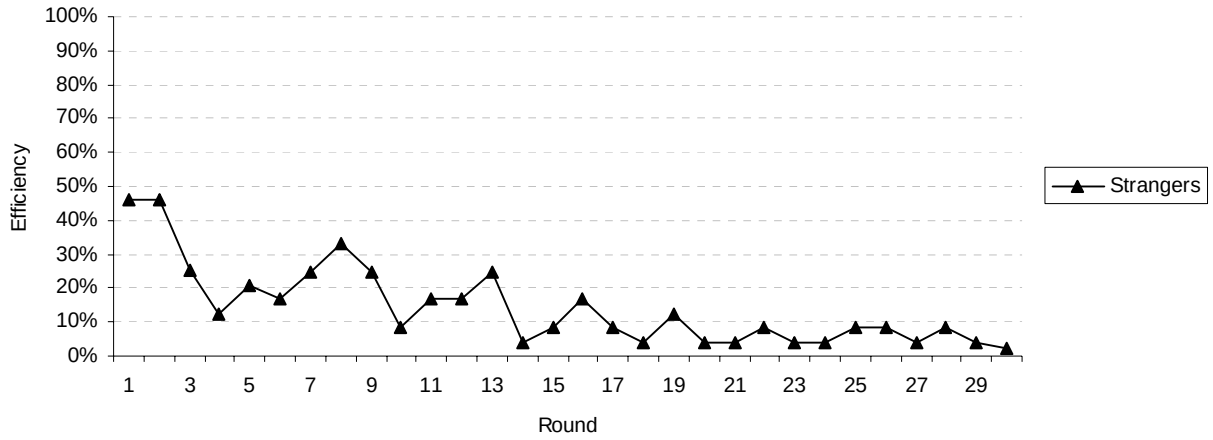

Figure 2. Strangers treatment, efficiency (% of completed trades) by round

[7] Payoffs, framing and context may all affect the exact numbers. However, based on our extensive research with various payoff parameters, different framings, contexts and experimental procedures (see reference list), we are confident that the qualitative results that we discuss in this paper are robust.

Aggregating over all rounds, there is almost the same amount of trustworthiness as in the one-shot version of the trust game in Figure 1. This reflects that the Strangers market does not create additional incentives to be trustworthy as compared to the one-shot game. On the buyer side, there is, on average, more trust in the Strangers market than in the one shot version of the game, possibly reflecting the hope that repeated action may support more cooperation. But the dynamics reveal that buyers respond to the fact that, on average, this expectation was disappointed: They started out by trusting quite a lot but trust quickly collapsed. In fact, the percentage of last round trust was only 0.04%, much less than in the one shot-game, indicating that buying in the one shot game is mainly due to inexperience.

Summing up, economic theory underestimates the degree of cooperation in one-shot encounters of anonymous traders; there is intrinsic trustworthiness. To the extent people cooperate, the need for a reputation system is diminished. However, there is not enough intrinsic motivation to stabilize positive reciprocity in an anonymous community without external enforcement in our setting. In this sense, economic theory is right: relying on solely intrinsic motivation will not, in the long run, lead to satisfactory cooperative behavior.

III. Extrinsic motivation: What is the gain of introducing a perfect reputation system?

Reputation systems provide an external enforcement device that may help to overcome the cooperation problems in anonymous communities. From an economic theory point of view the incentives created by reputation systems depend on the exact trading environment. Suppose, for the moment, that the buyer-seller encounter in Figure 1 is played repeatedly, with an infinite time horizon, and so with no expectation of a stopping round of play. In such a setting, even if all traders are selfish and rational, there exist equilibria in which the buyer always buys and the seller always ships. The equilibria can be supported by reciprocal trigger-strategies that call for a buyer, for instance, to trust as long as the seller has shipped when he or she has had opportunities to do so in the past. Once the seller defects, he will never be trusted again. If future payoffs are sufficiently important, this gives an incentive to the seller to be trustworthy all the time, and thus for the buyer to trust all the time (ex., Kandori, 1992). An interesting feature of this argument is that it is independent of whether the reputation system relies on direct or indirect reciprocity. The information available to the buyer about the seller is what is important; if the information is

sufficient in quantity and accuracy, the buyer can act on it just as well if the information were generated elsewhere or if it were generated from the buyer's own experience.

There are, however, two problems with this kind of simple equilibrium in our context. First, the trading horizon in online market platforms is typically finite. If either the buyer or the seller believes that there will be some upper boundary of items to be traded (so a finite horizon game), cooperation among selfish, rational traders will unravel (in the last round there is no trustworthiness, and so no trust and no trade, and for this reason no trade in the second to last round, etc.). Second, and perhaps more deeply, buying and shipping in the infinite game equilibrium does not really capture what we mean by trust and trustworthiness because, in equilibrium, there is no uncertainty about each others' behavior; in equilibrium, sellers *do* have a material incentive to ship, so there is no risk of being exploited, and so no trust involved in buying.[8]

Since this paper is concerned with trust (characterized by a risk of being exploited) in economic transactions (where traders typically trade a finite number of items), we think it more appropriate to study models of how trust can be built in finitely repeated games. In these theories, trust emerges when there is some (possibly small) amount of 'truly intrinsic' trustworthiness within the seller population (Wilson 1985). That there is intrinsic trustworthiness has been demonstrated, for instance, in our experimental studies of the one-shot trust game of Figure 1. In essence, in theory, the existence of some intrinsically trustworthy sellers gives all sellers an incentive to build a reputation as trustworthy, at least until towards the end of the game (at which point a good reputation is less valuable). Hence buyers can trust sellers, at least early on, because there is a high probability (albeit less than one in the last few rounds) that all sellers will *act* trustworthy.[9] It turns out that reputation building in this model, in the context of the buyer-seller encounter, is as in the infinite horizon models independent of whether the reputation system relies on direct or indirect reciprocity (Bolton and Ockenfels 2006), something we will come back to below.

[8] There are also more subtle equilibria in these models in which cooperation in any given round is uncertain, but this raises yet a third problem, that there are many equilibria in these models with outcomes ranging from full cooperation to no cooperation at all. In our view, trust is not satisfactorily described as an equilibrium selection problem.

[9] The mechanics of these equilibria are relatively complex and we will not delve into them here; see Bolton and Ockenfels (2006) for a theoretical and experimental treatment within a trust game environment.

So economic theory suggests that, in principle, reputation mechanisms of the sort we describe in the introduction can solve many of the trust problems associated with economic transactions. All the various models, finite and infinite horizon alike, suggest that sufficient information about past behavior is the key prerequisite to the emergence of trust, because it allows buyers to avoid sellers who are known as fraudulent and to buy only from sellers who have proven trustworthy in the past. Conditioning trust on the seller's history creates incentives for sellers to build up a reputation for being trustworthy, at least when the end of the market is not too close and maintaining a good reputation is still valuable. A reputation of being trustworthy can be developed and sustained even by completely rational and selfish sellers – as long as the probability of being matched with intrinsically trustworthy sellers is strictly positive. We know from the Strangers market that intrinsic trustworthiness alone is not enough to sustain a trading platform that has no reputation system, however, is demonstrated in our experimental Strangers market. So does a feedback system help promoting trust and trustworthiness as suggested by theory?

Bolton et al. (2004a) compared the Strangers market to a *Reputation market*, played over 30 rounds, in which, as before, a buyer never met the same seller more than once. However, in this market we introduced a reputation system that, similar to eBay's feedback forum, informs buyers about all past actions of their current seller (see Duffy and Ochs, 2003, and Bohnet and Huck, 2004, for related experimental work). This feedback information is always shared and reliable (because it is not given by the buyers themselves but by the experimenter), and sellers had no way to change their online identity. This way, the experiment studies the impact of feedback information on trading behavior when an ideal, frictionless reputation mechanism is available. And in (the finite horizon) theory, this should be sufficient information to enable trust and trustworthiness, and so successful trade.

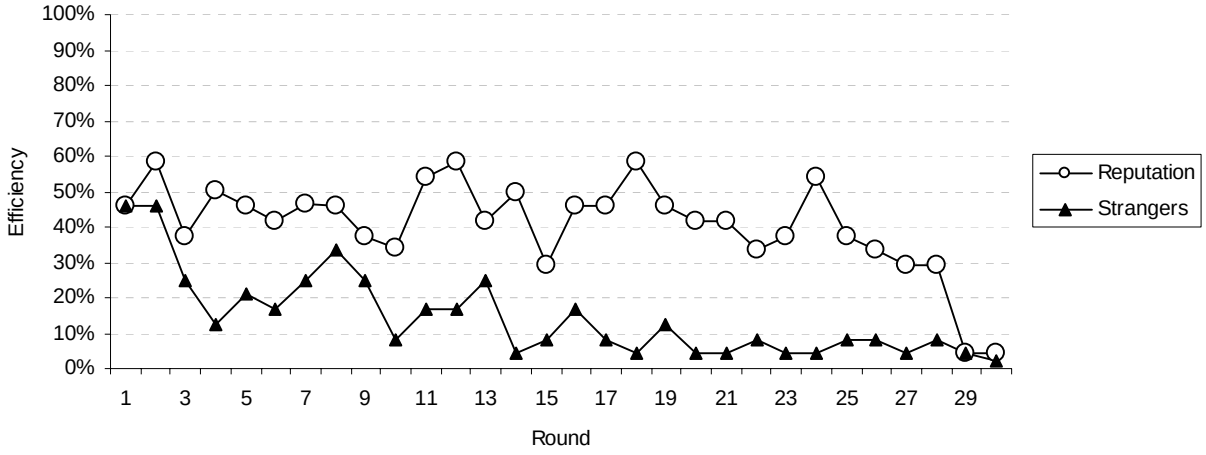

Figure 3: Reputation market, efficiency (% of completed trades) by round

Figure 3 shows the results of the Reputation market experiment and compares with the Strangers market results from Figure 2. On average, there is significantly more buying (56 vs. 37 percent; $p < .05$) and shipping (73 vs. 39 percent; $p < .01$) in the Reputation market than in the Strangers market. In fact, the shipping probability is slightly higher than the threshold of 70% for trust being profitable. As a consequence, the trade dynamics too look quite different than in the Strangers market; trading starts at about the same level as in the Strangers market but then the level remains stable until the very last rounds, when the strategic value of having a reputation for being trustworthy vanishes and virtually all cooperation collapses.

We conclude that introducing a perfect reputation system in a market with strangers has a strongly positive impact on trust, trustworthiness and trading efficiency. Both buyers and sellers respond strategically to the information provided. At the same time, however, the experiment demonstrates that there are serious limits of perfect reputation systems in promoting cooperation. The realized surplus as a percentage of potential surplus is only 41 percent. The gain from introducing a perfect system into a Strangers market (described in the last section) as a percentage of the maximal potential gain is 41 – 14 = 27 percent, which is well below what would be expected theoretically (see Bolton and Ockenfels 2006). Obviously, trader behavior is different from what we expect from theory, in a way that limits the effectiveness of reputation systems.

IV. What behavior limits the effectiveness of reputation systems?

We have seen that even though reputation systems can build on intrinsic motivations to cooperate, their effectiveness is less than what can be expected from theory based on purely selfish traders. What is the source of these limitations?

Since the feedback system in the experiment is perfect with regard to the information it delivers, we need look at departures from fully rational behavior for answers. There is evidence in the data that forward looking behavior is more limited than theory anticipates. Perhaps the strongest evidence for this is that out-of-equilibrium behavior is observed in the early rounds of play. For instance, the sellers' payoffs are strongly *positively* correlated with the overall number of shippings; the Spearman rank correlation is 0.504 (p = 0.000). Shipping early is not only trustworthy and fair, it also pays. However, many sellers have difficulties to understand the future benefits of being nice. About 40% of the sellers in the Reputation market who receive an order in the first round of the market fail to ship.

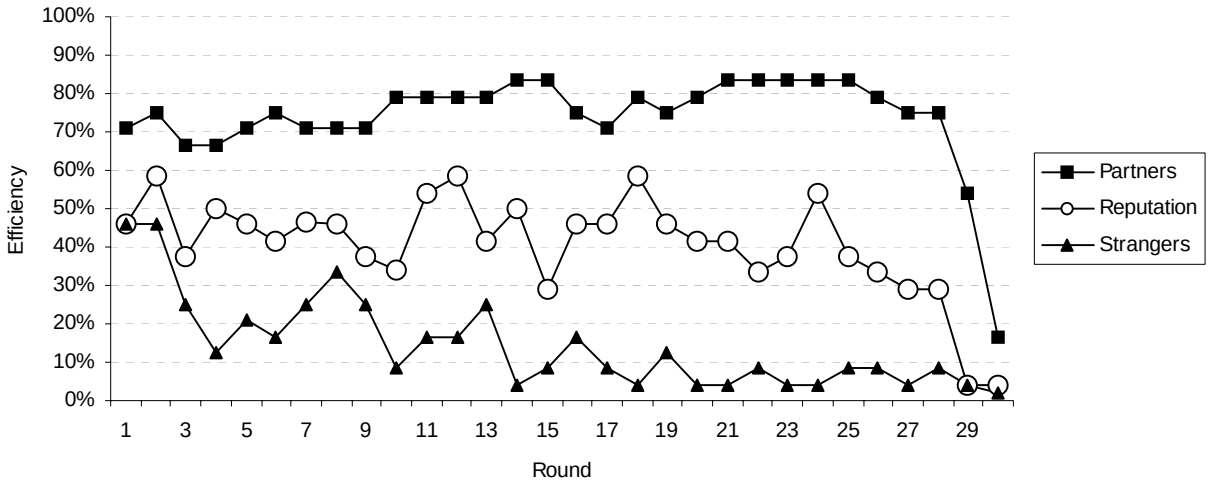

Figure 4: Partners market, efficiency (% of completed trades) by round

There is also evidence that traders learn from looking back, a kind of learning the equilibrium model does not anticipate. For example, Reputation market sellers are actually more inclined to ship in the middle rounds of the market than at the beginning. And in the Strangers market, the 65% of buyers who start out trusting quickly learn that they should not. This behavior is consistent with "low rationality" adaptive learning models that suggest that people

come to strategic games with rough priors and adjust these priors according to the payoff reinforcement they get from experimenting with various strategies (ex., Erev and Roth 2000).

There is also additional evidence to suggest that bounded rationality is not the entire story for why Reputation market trading performance falls short of what theory leads us to expect. Bolton et al. (2004a) also included a "Partners market." The only respect in which this market differed from the Reputation market is that, in the Partners market, the same buyer was matched with the same seller for the entire market. Recall that theory suggests that there should be no difference in the performance of the two markets: in both cases, buyers should be able to play tit-for-tat strategies to keep sellers trustworthy. Nevertheless, Figure 4 demonstrates there is a substantial difference between them. Overall, trading (efficiency) levels in the Partners markets, 74 percent, is significantly higher than in the Reputation market ($p < 0.025$).

The amount of trading in the Partners markets is still substantially less than perfect, indicating that bounded rationality explanations given above still apply. Still, trade efficiency is greater than in the Reputation markets, and this suggests that some other things beside bounded rationality are at play. In Bolton et al. (2004a) we argue that the flow of information in the Reputation markets creates *information externalities* in that, out-of-equilibrium, the incentives to invest in the two markets are different. Specifically, there is a public goods problem in the Reputation market not present in the Partners market. Buyers do not benefit from the reputation information they themselves produce. As a consequence, Reputation market buyers under invest in the production of reputation information relative to Partners markets. [10]

So boundedly rational trading is off the equilibrium path, and the resulting *out-of-equilibrium incentives* may in turn affect traders' behavior. A second observation in this regard is that reputation information (even in a system with comprehensive and reliable feedback information) need be interpreted as a noisy signal, because the predictive value of reputation information suffers from the noise generated by the behavior of 'real' traders. This has consequences. As we have experimentally shown in a recent paper (Bolton, Loebbecke and Ockenfels 2006), market *competition* tends to increase the effectiveness of reputation systems in

environments with noisy behaviors. The reason is that with competition buyers can discriminate between sellers on the basis of the reputation information provided by the reputation system, creating *stronger* incentives for sellers to behave consistently trustworthy over time. The experiments in Bolton, Loebbecke and Ockenfels (2006) involve *matching competition* (each buyer gets to choose between two sellers; prices are fixed) and *price competition* (the two sellers compete on prices) to both the Reputation and Partners markets described in the previous sections. Our experiments show that seller competition in (perfect) Reputation markets typically enhances trust and trustworthiness, and always increases total gains-from-trade. We also find that information about reputation trumps pricing in the sense that traders usually do not conduct business with someone having a bad reputation not even for a substantial price discount. Thus price competition does not significantly undermine the sellers' incentives to be trustworthy. Finally, we find that a reliable reputation system can largely reduce the advantage of Partners markets (compared to Reputation markets) in promoting trust and trustworthiness described in earlier sections *if the market is sufficiently competitive*. So, one important overall conclusion from the study by Bolton, Loebbecke and Ockenfels (2006) is that, in a world with noisy traders and well-functioning reputation systems, encouraging greater market competition may be a powerful tool for increasing cooperation and trade efficiency.

V. A note on complex reputation measures

In the studies we have discussed so far, reputation is equivalent with information about the sequence of a seller's shipping decisions. Theoretically, this measure is sufficient to sustain cooperation in communities where reputation is built by indirect reciprocity. In fact, simple and stable cooperation in our settings can theoretically be reached with just information about a seller's *last* shipping decision, because this information is all one needs to employ reciprocal tit-for-tat strategies.

[10] It turns out that, while there are no information externalities in the incomplete information model of reputation building on the equilibrium path of the buyer-seller encounter, there are information externalities for the incomplete information model of reputation building of other market games such as Selten's chain store game; see Bolton and Ockenfels (2006) for a discussion and a formal experiment that shows that the phenomenon is more robust than theory suggests it should be.

However, in many markets reputation must be built on much more complex and multi-dimensional facets of the seller history. When assessing a seller's trustworthiness, buyers need to take into account, for instance, technical and cultural communication problems, the possibility of incomplete or manipulative feedback, the reliability of the "postal service" and so on. Even 'perfect' reputation mechanisms, which deliver all relevant information in order to promote cooperative interaction, may become quite complex, so that real traders experience information overload.

The information requirements in two-sided reputation systems are necessarily much more demanding than one-sided systems. For example, consider a system, in which buyers rate sellers and sellers rate buyers in order to mitigate moral hazard incentives on both market sides.[11] Now suppose that a buyer receives reliable information that the seller did not send the object to the last buyer. Does this imply that the seller is not trustworthy? No. It could be that the seller did not ship because his or her last buyer never sent the payment. Let's think this one step further. Would it then be enough for our current buyer to know whether the current seller's last buyer paid? Again, the answer is no. Whether the last buyer's action can be interpreted as trustworthy depends on the history of play of his or her earlier transaction partners. In principle, the entire history of both trading partners as well as their trading partners, and their trading partners, and so on, may be required to construct a system that has sufficient information of the sort we tested in the one-way settings discussed earlier.

Clearly, this information is difficult to process, even when comprehensive and perfectly reliable. There is a way to avoid the processing problems, though. The relevant information can be captured in a single reputation rating, which does not directly reveal past behavior but rather evaluates these behaviors according to all traders' histories and with respect to a trading norm. This rating can, in theory, be easily processed. On the other hand, however, the information content is less comprehensible due to the rather complex information aggregation processes behind the rating.

Let us illustrate the issues with the help of the simple 'image scoring game' (Nowak and Sigmund 1998). As with the markets we studied in the last sections, the image scoring game

conceives of the group interacting over a series of rounds. Again, in each round, people are paired off at random. One person in the pair, designated the title of 'mover,' is given the opportunity to give a favor to the other, designated the title of 'receiver.' These designations are assigned randomly, so over many rounds, each player is a mover about half the time and a receiver the other half. Giving a favor would cost the mover c and benefit the receiver $b > c > 0$. Figure 5 illustrates the situation.

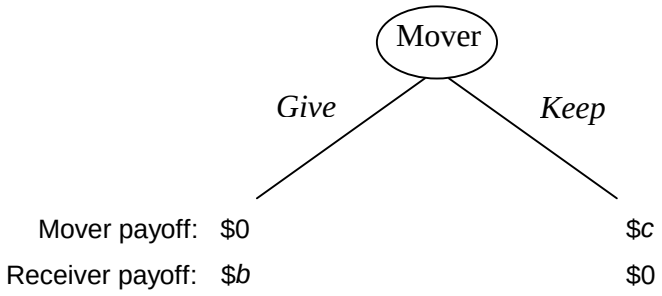

Figure 5. When mover meets receiver in the image scoring game

The efficient outcome in this game, the outcome that maximizes the total social benefits, is for everyone to give when they are the mover. While keeping maximizes short-run payoffs, reputation can help, by providing the information necessary to reward those who give with giving and punishing those who do not with keeping. This kind of reciprocity is not unlike the trust game context we discussed in earlier sections. However, even though the game looks much simpler than the game in Figure 1, the basic reputation issue is more complicated than before. To see why, consider the kind of reciprocity that works in the trust game markets based on the game in Figure 1. The mover gives if he knows the receiver played give the last time as a mover, and keeps if the receiver last played keep. Suppose now that you are the mover matched with someone who last played *keep* as a mover. If you play *keep* as the reciprocity strategy stipulates, then the next time you are the receiver, you can expect the mover to play *keep* on you (if others too play the reciprocity strategy). Consequently, you make more money playing *give* (lose c now, pays b later) than playing *keep* (gain c now, pays 0 later). The problem is that if enough

[11] This is not much unlike eBay, where both transaction partners can rate each other. However, because eBay transactions are typically done sequentially (first buyers send money, then sellers ship the object), the moral hazard
(continued on next page)

people decide to *give* to keepers then it pays to be a keeper. And if it pays to have a bad reputation, then why have a good one?

So, this kind of *first order information* about what the opponent did last time as a mover is, theoretically, not enough to stabilize cooperation. If we add *second order information*, the receiver's reputation would include not only what he did last time as a mover, but also what the receiver he faced did last time as a mover. For example, the reputation might reveal that the receiver "last played keep with a player who last played give." This amount of recursive information pushes the unraveling problem back by a step. To see this, consider a mover who, for the first time, encounters a receiver who played keep on a giver. To support his punishment, keeping on a keeper would have to be rewarded, meaning that there needs to be giving to someone who gives to a keeper – which is not consistent with self-interest since keeping on a keeper pays more. So now players would have to think *two steps* ahead, and be confident others do so as well, before cooperation would unravel.

Of course, if traders think three steps ahead, this is still not enough. To stabilize cooperation in a population of rational traders, one would need the *entire* transaction histories of basically *all* traders. For this reason, some theorists have cautioned that indirect reciprocal systems might not be stable outside of very small groups where the information demands are relatively modest.

Boundedly rational traders, however, often do not think many steps ahead (see the previous section and, e.g., Nagel 1995), and people's ability to do backwards induction is rather limited. In fact, in an experimental study of the image scoring game (see Bolton, Katok and Ockenfels 2005 for the details)[12] we find that first order information significantly increases cooperation rates above the level in a market without any reputation information. Second order information again significantly increases cooperation rates, reflecting that trader do some of the backward induction, but do not think through the whole problem. However, both markets with strangers matching perform dramatically worse than the corresponding Partners market.

---

problem is mostly on the seller side.

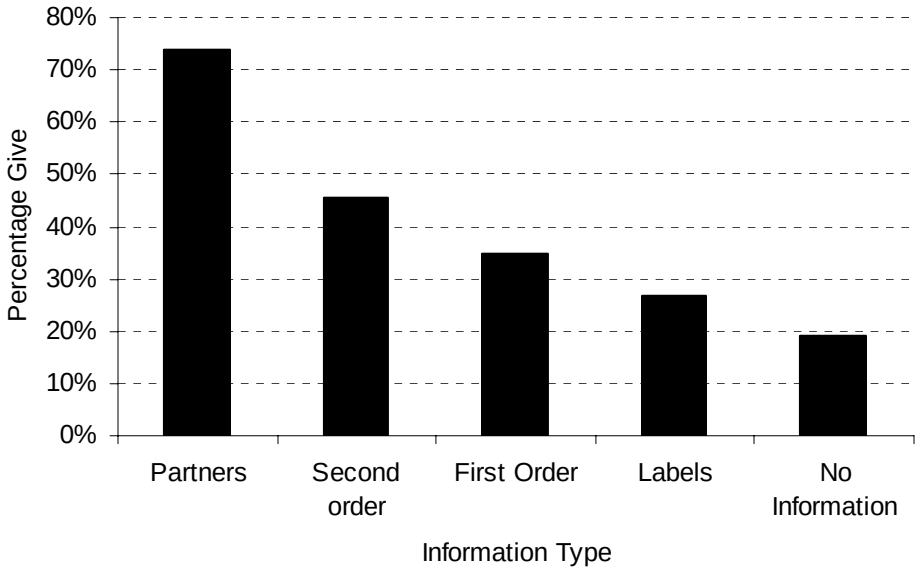

Figure 6. Giving levels (averaged over all rounds) in experimental image scoring games

How can the gap of the effectiveness of reputation systems between partners and strangers matching markets be closed? We think it is unlikely that higher order information may be of considerable help, because second order information is already difficult to communicate and to process. One way could be to aggregate *all* the relevant information into a single reputation score, so that traders might then apply a simple reciprocity strategy in a way that cannot be cheated on (Kandori 1992).

In our experiment (Bolton, Katok and Ockenfels 2005) we proposed the following reputation score along these lines: We labeled each player in each round as a member of either the 'matcher' club or the 'non-matcher' club according to the following rule: In the first round, everybody is a matcher. In every round after that, a player's label is updated as follows: (i) if the player gave to a matcher the last time he was mover, he is a matcher; (ii) if the player kept on a

[12] . Subjects were Penn State University students, mostly undergraduates from various fields of study, and recruited by fliers posted around campus. In total, there were 192 participants. We ran two image scoring games for each information condition, each game with 16 subjects playing for 14 rounds. Each round, subjects were anonymously paired, interfacing with one another via computers. The value of a gift, *b*, was $1.25 and the cost of giving, *c*, was *(continued on next page)*

non-matcher, he is a matcher; (iii) if the player did anything else, he is a non-matcher. Now consider a reciprocity strategy that stipulates giving to a matcher and keeping on a non-matcher. If everyone follows this rule, then everyone will stay a matcher and there will be 100 percent cooperation. Moreover, you cannot benefit by cheating. If you keep on a matcher, you become a non-matcher which lines you up to be punished since the next time you are matched with a mover, he will keep on you. And punishment is now with impunity: keeping on a non-matcher allows a mover to maintain matching status – he won't be punished for doing the right thing.

When *all* information is processed in this way, the reciprocal strategy yields stable cooperation – at least in theory. To our surprise, however, the experimental data does not confirm at all the prediction. The bar called "labels" in Figure 6 shows the average giving rate in this setting. The information that should stabilize the cooperation rate in fact significantly *reduces* the cooperation rate compared to the other settings, which involve theoretically insufficient reputation information. It appears that real traders have difficulties with reputational reports that *filter* actions, and respond more favorably to reputational reports about recent past *actions*. The dilemma is that this information, when being complete, cannot be processed by boundedly rational traders.

VI. Conclusions

What we learn from the experimental and theoretical work is that it is the *interplay* of institutions with bounded rational behavior that drives the results. No doubt, institutions matter. But behavioral aspects of reputation building matter too. As a result, standard economic models based on full rationality and narrow self interest tend to overestimate the difficulties of promoting trust in one shot situations, and underestimate the difficulties in ongoing interaction in communities of strangers.

Because the laboratory reputation systems we study here are perfect, the limits of their effectiveness cannot be the result of institutional defects but must be due to 'behavioral defects'. That is, the restraints that we observe are rooted in boundedly rational behavior. There are basically two types of 'noisiness' in the behavior that significantly affect the functioning of the

---

$0.75. Subjects knew that they would be in each role, mover or receiver, for half the trials (7 times) and roles would generally rotate between rounds.

institutions. For one, bounded rationality can directly affect trust and trustworthiness via non-rational choices. Besides difficulties that arise when handling complex reputation measures, we observe that 'real' traders have difficulties to cope with reputation building *dynamics*. Many traders fail to sufficiently look forward and to fully take into account the future consequences of current behavior. Other behaviors are characterized by too much backward looking and simple, adaptive learning patterns. Second, noisy behavior moves the reputation building dynamics off the equilibrium path and thus changes (out-of-equilibrium) incentives in ways that systematically affect strategic reputation building. We observe, for instance, that noisy behavior creates information externalities so that the flow of reputation information through the community becomes critical for the effectiveness of reputation systems. Also, when trading dynamics are out-of-equilibrium, seller competition becomes a powerful support for reputation systems.

We think that only a combination of complementary field, laboratory and thought experiments can reveal the full story behind reputation systems. Field studies strive for external validity and require a careful look at institutions. They make it difficult, if not impossible, however, to separate institutional from behavioral influences, to measure the impact of either aspect on the effectiveness of reputation systems, and to measure the overall impact of a reputation system. Thought experiments (e.g., equilibrium theory) help to understand how behavior and institutions interact, reveal basic incentive structures and allow generalizing from empirical observations. But it is risky not to complement thought experiments with data, because it is known that theory can sometimes yield dramatically wrong conclusions, especially when it comes to social interaction (e.g., Bolton and Ockenfels 2000). Thought experiments also tend to neglect institutional details, which can turn out to be very critical, both in the equilibrium analysis and in 'reality' (e.g., Klemperer 2004). Laboratory experiments can separate and measure the different impacts, and the interplay between institutional and behavioral influences. Combined with field and thought experiments, they are a powerful tool to analyze the effectiveness of existing and newly designed reputation systems.

References

Bohnet, I. and S. Huck (2004), "Repetition and Reputation: Implications for Trust and Trustworthiness When Institutions Change." *American Economic Review (Papers and Proceedings)*, 94(2), 362-366.

Bolton, Gary E, Elena Katok, and Axel Ockenfels (2004a). "How Effective are Online Reputation Mechanisms? An Experimental Investigation." *Management Science*, 50(11), 1587-1602.

Bolton, Gary E, Elena Katok and Axel Ockenfels (2004b). "Trust among Internet Traders. A Behavioral Economics Approach." *Analyse und Kritik*, 26, 185-202.

Bolton, Gary E, Elena Katok, and Axel Ockenfels (2005). "Cooperation among Strangers with Limited Information about Reputation." *Journal of Public Economics*, 89, 1457-1468.

Bolton, Gary E, Claudia Loebbeke, and Axel Ockenfels (2006). "How Social Reputation Networks interact with Competition in Anonymous Online Trading: An Experimental Study". Working paper, University of Cologne.

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

Bolton, Gary E, and Axel Ockenfels (2006). "Reputation, Matching and Information Externalities: Theory and Evidence." Working paper, University of Cologne.

Dellarocas, Chris (2003), "The Digitization of Word-of-Mouth: Promise and Challenges of Online Reputation Mechanisms," *Management Science*, 49, 1407–1424.

Dellarocas, Chris (forthcoming), "Reputation Mechanisms," in: Handbook on Economics and Information Systems (T. Hendershott, ed.), Elsevier Publishing.

Dellarocas, C., M. Fan, and C. Wood (2003), "Self-Interest, Reciprocity, and Participation in Online Reputation Systems," *2003 Workshop in Information Systems and Economics (WISE)*, Seattle, WA.

Duffy, J. and J. Ochs (2003), "Cooperative Behavior and the Frequency of Social Interaction," Working paper, University of Pittsburgh.

Dufwenberg, Martin, and Georg Kirchsteiger (2004), "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47, 268-98.

Fehr, E., and Schmidt, K. (1999), "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817-868.

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

Güth, Werner, Friederike Mengel, and Axel Ockenfels (2006). "An Evolutionary Analysis of Buyer Insurance and Seller Reputation in Online Markets." Working paper, University of Cologne.

Kandori, M. (1992). "Social Norms and Community Enforcement." *Review of Economic Studies*, 59, 63-80.

Klemperer, Paul (2004), "Auctions: Theory and Practice," Princeton University Press.

Ledyard, J. (1995), "Public goods: a survey of experimental research," in: Kagel, J.H., Roth, A.E. (Eds.), Handbook of Experimental Economics. Princeton: Princeton University Press, 111–194.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser (2005), "Eliciting Honest Feedback: The Peer Prediction Method," *Management Science*, 51(9), 1359-1373.

Nagel, Rosemarie (1995), "Unraveling in Guessing Games: An Experimental Study". *American Economic Review*, 85(5), 1313-1326.

Ockenfels, A. (2003), "Reputationsmechanismen auf Internet-Marktplattformen." *Zeitschrift für Betriebswirtschaft*, 73(3), 295-315.

Ockenfels, A., und J. Weimann (1999), "Types and Patterns – An Experimental East-West-German Comparison of Cooperation and Solidarity". *Journal of Public Economics* 71(2), 275-287.

Rabin, Matthew (1993), "Incorporating Fairness into Game Theory and Economics". *American Economic Review* 83, 1281-1302.

Rashid, Al Mamunur, Ling, Kimberly, Tassone, Regina D, Resnick, Paul, Kraut, Robert, and Riedl, John 2006). Motivating Participation by Displaying the Value of Contribution. In *Proceedings of ACM CHI 2006*, 955-958.

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

Resnick, Paul, Richard Zeckhauser, John Swanson, and Kate Lockwood (2006), "The Value of Reputation on eBay: A Controlled Experiment," *Experimental Economics*, 9(2) 79-101.

Wilson, Robert (1985), "Reputations in Games and Markets," in Alvin Roth ed., *Game Theoretic Models of Bargaining*, Cambridge University Press, 27-62.