The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions*

John A. List
University of Maryland and NBER

17 June 2004

Abstract

The role of the market in mitigating and mediating various forms of behavior is perhaps the central issue facing behavioral economics today. This study designs a field experiment that is explicitly linked to a controlled laboratory experiment to examine whether, and to what extent, social preferences influence outcomes in actual market transactions. While agents drawn from a well-functioning marketplace reveal strong reciprocity motives in tightly controlled laboratory experiments, when observed in environments that more closely resemble their naturally occurring settings, their behavior approaches what is predicted by self-interest theory. In the limit, much of the observed behavior in the marketplace that is consistent with social preferences is due to reputational concerns: suppliers who expect to have future interactions with buyers provide higher product quality only when the buyer can verify quality via a third-party certifier. There is, however, empirical evidence suggesting that social preferences influence outcomes in long-term relationships. In these transactions, the reputation effect is roughly twice as large as the social preference effect.

JEL: C93 (Field Experiments)
Key words: social preferences, field experiment

Correspondence to: John A. List, Professor, The University of Maryland, 2200 Symons Hall, College Park, MD 20742-5535, email: jlist@arec.umd.edu; website: http://www.arec.umd.edu/jlist/.

*Orley Ashenfelter, Raymond Battalio, Roland Benabou, Daniel Benjamin, Gary Charness, Edward Glaeser, Uri Gneezy, Glenn Harrison, Daniel Kahneman, Liesl Koch, David Laibson, Matthew Rabin, and Al Roth provided remarks on an earlier version of this study that improved the paper. Seminar participants at Harvard University, Princeton University, University of Texas at Austin, and Texas A&M provided comments that helped to shape the paper. Thanks to Michael Price for research assistance.

I. Introduction

More than two decades ago, George Stigler (1981) wrote that when "self-interest and ethical values with wide verbal allegiance are in conflict, much of the time, most of the time in fact, self-interest theory….will win." While this is the conventional wisdom among economists, an influential set of laboratory experiments on "gift exchange" has provided strong evidence that Stigler's position is often not valid (see, e.g., Camerer and Weigelt, 1988; Fehr et al., 1993; Berg et al., 1995). This literature is complemented by an entire body of research relating to theoretical explanation of social preferences (for models of reciprocity see Rabin, 1993, Dufwenberg and Kirchsteiger, 1999, Falk and Fischbacher, 1999, and Charness and Rabin, 2002; for models of inequity aversion see Fehr and Schmidt, 1999, and Bolton and Ockenfels, 2000; on altruism see Andreoni and Miller, 2002) and experimental studies designed to explore further the nature of social preferences and the robustness of the gift exchange results (e.g., Charness, 1996; Fehr et al., 1997; Fehr and Falk, 1999; Charness and Rabin, 2002; Gächter and Falk, 2002; Hannan et al., 2004; Brown et al., 2004; Fehr and List, 2004).[1]

The general results, which are consistent with the notion that people behave in a reciprocal manner even when the behavior is costly and yields neither present nor future material rewards, have attracted much attention, as many have argued that they are relevant beyond the context inherent in the laboratory experiments. For example, many view the experimental results as providing key support for the labor market predictions in Akerlof (1982) and Akerlof and Yellen, (1988; 1990), whereby higher than market-clearing wages and involuntary unemployment are potential outcomes of fairness considerations in the workplace.[2] Indeed, Fehr et al. (1993, p. 437)

[1] Fehr and Gächter (2000) provide an excellent overview. The interested reader should also see the related literature on "lemons" markets (e.g., Miller and Plott, 1985; Holt and Sherman, 1990; Lynch et al., 1991).

[2] This conjecture is typically termed the "fair wage-effort" hypothesis. Alternatively, note that the "efficiency wage theory" surmises that wages above market-clearing levels occur because these wage profiles induce workers to be motivated in an effort to avoid being fired, which economizes on firm-level monitoring (see, e.g., Katz, 1986).

note that their results "provide…experimental support for the fair wage-effort theory of involuntary unemployment." Of course, social preferences may be important in many other strategic situations as well (for overviews see, e.g., Camerer, 2002, and Sobel, 2002), and therefore such results have broad implications for economists and non-economists alike.[3] Despite these advances and the topic's importance, it is fair to say that little is known about whether, and to what extent, social preferences influence economic interactions in naturally occurring markets.[4]

The major goals of this study are to explore the nature of such preferences among real market players in naturally occurring environments and to provide a framework with which to disentangle social preferences and reputation effects. Measuring and disentangling social preferences and reputation effects is important in both a positive and normative sense, as optimal contracting and proposed government intervention in principal-agent settings, appropriate designing of collective choice mechanisms, and theory-testing all depend critically on proper measurement of these effects. To complete these tasks, I use several distinct experimental treatments that explicitly link laboratory experiments with field experiments. The field experimental setting mirrors the laboratory gift exchange experiments and resembles many types of good or service markets: after receiving a price offer, sellers determine the good's quality, which cannot be perfectly measured by buyers. This unique aspect of the experimental design also permits me to examine whether individual behavior in laboratory experiments provides a reliable indicator of behavior in the field.

[3] The results have also been used explicitly to test game theoretic predictions. In this study, I define "social preferences" to be preferences that are measured over one's own and others' material payoffs. In this respect, I am not interested in pinpointing whether the behavior consistent with social preferences is altruism, reciprocity, inequality-aversion, or based on another motive. For a parsing of trust and reciprocity in a laboratory experiment see Cox (2004).

[4] There is some survey evidence reported from interviews with managers that social preference considerations are important in the workplace (Blinder and Choi, 1990; Bewley, 1995). Furthermore, in a novel paper exploring the role of fairness in the marketplace, Kahneman et al. (1986) report results from telephone surveys of residents of two Canadian metropolitan areas (Toronto and Vancouver). Their data are neatly explained by a "dual entitlement" theory: previous transactions establish a reference level of consumer and producer surplus, and fairness considerations arise from outcomes relative to these "entitlements."

Treatment I has subjects drawn from a well-functioning marketplace—the sportscard market—participating in gift-exchange laboratory experiments that closely follow the received literature. In these experiments, consumers are placed in the role of buyers and dealers are placed in the role of sellers. Experimental results are broadly consistent with the literature that uses students as subjects: the evidence suggests that social preferences have an important influence on economic outcomes. This finding provides a nice validity check of the extant laboratory results on social preferences, as it suggests that the major results can be replicated with real economic players from a much different population.

Treatment II recognizes that the (relatively) context-free setting in Treatment I is devoid of potentially important elements of the exchange process and therefore may suppress important psychological effects. Thus, in Treatment II, I draw subjects from the same subject pool, but instead of using (relatively) context-free instructions, I add context that closely resembles the subjects' naturally occurring environment. For example, the generic induced value setting in Treatment I is now augmented by having buyers make an offer to a seller to buy one 1990 *Leaf* Frank Thomas baseball card, and sellers subsequently choosing the quality of the baseball card if they accept the buyer's offer. If one ignores the artificiality invoked by the laboratory experimental setting, this particular treatment provides an environment closely related to the actual decision-making process in the marketplace from which these subjects are drawn. This simple design change yields behavioral differences, but gift exchange in this setting remains alive and well, both statistically and economically.

Treatments III$20 and III$65 represent the naturally occurring analogues to Treatment II. In Treatment III$20, subjects approach dealers (who are unaware that they are taking part in an experiment) who have several 1990 *Leaf* Frank Thomas sportscards on hand and offer $20 for a

"Thomas card that would grade at least PSA 9."[5] The two design parameters ($20 and the requested product quality) were chosen to closely match the average price and requested quality observed in Treatment II ($20 and PSA 9). Treatment III$65 is identical in structure: buyers approach dealers on the floor of a sportscard show but now offer $65 for a "Thomas card that would grade at PSA 10." Since quality is difficult to detect in this market for untrained consumers, if social preferences play a role in this case the card's grade and the price offer should be positively correlated. Once the buying agents had purchased each of the cards from the dealers in Treatment III, I had every card professionally graded. I do find such a correlation between the prices and grades received, but only among dealers who are "locals"; among dealers who are likely to have little future interaction with the buying agents, no such relationship emerges.

This result suggests that reputation effects are important in this market, but such findings may be due to several factors, including sample selection (i.e., local dealers have social preferences and non-local dealers do not). A final set of treatments—denoted Treatments IV-NG, IV-AG, and IV-G—provide insights into what is driving these behavioral differences by examining outcomes in an identical experiment for collector tickets and ticket stubs. Tickets and ticket stubs provide a unique test because no third-party verification service existed to grade tickets until recently (June 2003). In this sense, by comparing outcomes before third-party verification was possible with outcomes after grading services were available, I have a unique opportunity to examine not only the nature of market exchanges with and without third-party enforcement, but I am also able to explore the role of social preferences in such settings. Brown et al. (2004, p. 7) summarize the attractiveness of such treatments when they motivate their laboratory experiments by noting "The ideal data set for studying the effects of the absence of third party enforceability on market

[5] PSA (Professional Sports Authenticator) is the major grading company in the industry and uses a 1-10 scale, with 10 representing the highest quality. See below for more detailed remarks on sportscard grading.

interactions…is based on a truly exogenous ceteris paribus variation in the degree of third party enforceability……The problem is, however, that it seems almost impossible to find or generate field data that approximates this ideal data set." This is exactly what Treatment IV offers, and to my best knowledge such exogeneity has not heretofore been achieved in this literature.

Treatment IV-NG (denoting no-grading available) is similar to Treatment III: at sportscard shows between October 2002 and March 2003, subjects approached dealers and offered $10 ($30) for a "ticket that would grade at least PSA 9 (10) if professional grading was available."[6] Unlike Treatment III data, the empirical results in this case provide little evidence consistent with social preferences: ticket quality is not correlated with price and local and non-local dealers provide similar quality levels. One could reason that dealers had little idea how to grade tickets since they had never been graded to date (even though many dealers made quality claims), and therefore the inability for Treatment IV-NG to reject the homogeneity null is perfectly consistent with informational problems.

This potential problem is rectified in Treatment IV-AG (denoting announcement of grading), which was administered at sportscard shows after PSA announced they would begin grading ticket stubs (April 2003) but before they released their grading criteria (June 2003). Purchasing identical tickets and using identical protocol to Treatment IV-NG, I find that during this time period gift exchange is prevalent among local dealers but not among non-locals: quality and price are correlated for tickets sold by locals but no correlation is present in ticket sales among non-locals. This result is entirely consistent with the empirical findings in Treatment III using sportscards.

Completing the experimental design is Treatment IV-G (denoting grading available), which is identical to Treatments IV-NG and IV-AG, but was completed post-June 2003. Insights gained

[6] The price adjustment was made to account for differences in card versus ticket values.

from Treatment IV-AG and IV-G data are quite similar, which stands to reason because PSA's ticket grading criteria is very similar to its scheme for grading sportscards—which has proven quite popular, as PSA has graded more than 7 million sportscards to date.

In summary, several insights follow. First, even though the data collected from one-shot laboratory experiments suggest that social preferences are quite important among this lot of subjects, parallel treatments in the field suggest that such effects have minimal influence in naturally occurring transactions. In this sense, dealer behavior in the marketplace approaches what is predicted by self-interest theory. Yet there is evidence that relationship length is important in market outcomes: in those cases where the seller and buyer have had considerable previous interaction, gift exchange is evident even in the absence of third-party verification. The measured social preference effect in such transactions is roughly half the size of the estimated reputation effect. Second, empirical results suggest that third-party enforcement of contracts is important, even when the market is populated by individuals with social preferences. This result follows from the (ubiquitous) increased level of delivered product quality when third-party enforcement was available. While theory has progressed substantially during the last two decades, the overall set of results provides new challenges for theorists and empiricists alike, as they suggest that crucial gaps in our knowledge about the effects of contracts and incentives exist.[7]

The remainder of this study is organized as follows. Section II describes the experimental design and summarizes the institutional details of the market. Section III provides a discussion of the empirical results. Section IV concludes.

II. Experimental Design and Institutional Details

The experimental investigation begins with an examination of behavior in standard laboratory gift exchange games. Treatment I-R (R denotes laboratory replication—see Table 1 for a

[7] Prendergast (1999) and Chiappori and Salanie (2003) provide excellent summaries.

summary of the experimental design) makes use of the general gift exchange experimental design. One session was run in this treatment. In this session, each participant's experience typically followed four steps: (1) consideration of the invitation to participate in an experiment, (2) learning the experimental rules, (3) actual participation, and (4) conclusion of the experiment and exit interview. In Step 1, the monitor approached dealers on the floor of a sportscard show and inquired about their interest in participating in an economics experiment that would take about an hour. If the dealer agreed, the monitor summarized the meeting time and place. Since most dealers are accompanied by at least one other employee, it was not difficult to obtain agreement after it was explained that they could earn money during the experiment. A similar approach was used to recruit consumers (non-dealers).

Subjects met in a large room adjacent to the sportscard show floor: dealers entered on one side of the room and non-dealers on the other side, and a divider was in place to ensure that identities were not revealed. The session consisted of five periods, with five dealers acting as sellers and five non-dealers acting as buyers. Each participant received a copy of the instructions, and to ensure common information the monitor read the instructions aloud as the subjects followed along.[8] The instructions noted that in each of the five periods each buyer would be paired with a different seller. In every period, the buyer determines an integer value (denoted p for price) to send to the seller, and requests a specific quality of the good (denoted q_r for quality request). Only the seller who is paired with the buyer is aware of these two choices. After the buyer makes these private decisions on the decision sheet, the monitor collects the sheets and walks them to the seller partners. Sellers then choose a quality level (denoted q for quality chosen), with an associated cost of quality (denoted c(q)—see Appendix A for the cost of product quality parameters) that is

[8] Appendix A contains a copy of the instructions.

increasing monotonically with product quality. The product quality choice is revealed only to the buyer partner (all choices are revealed to the monitor, of course).

Individual p and q choices combine to determine monetary payoffs for the pair according to the following payoff functions:

Seller payoff: $\prod_s = p - c(q)$ (1)

Buyer payoff: $\prod_b = (v - p)q$ $v = \$80$, $p \in [\$5, \$80]$, $q \in [.1,1]$

All payoff information was common information, and before beginning the experiment several hypothetical exercises were completed to ensure that everyone understood the instructions and payoff functions. Subjects were also aware that one of the five periods would be selected randomly and that that particular period would determine payoffs. After the fifth period, subjects were paid in private after they completed the survey contained in Appendix B.

These parameter values yield a standard prediction under the assumption of common knowledge, self-interest theory, and appropriate backward induction. Since product quality is costly, sellers will choose the minimum level ($q_{min} = 0.1$). A buyer's best response is to choose p_{min}, which is p = \$5. Thus, the subgame perfect equilibrium outcome is q* = 0.1 and p* = 5, with associated profits of $\prod_s = \$5$ and $\prod_b = \$7.5$, much less than more efficient profit levels (i.e., p = 30 and q = 0.5 yields $\prod_s = \$24$ and $\prod_b = \$25$). Previous experimental efforts have found that typically q > q* and p > p*, leading to an interpretation that reciprocity is important in economic interactions. More generally, this result suggests that people respond to acts that are perceived as kind in a kind manner.

Moving to column 2 in Table 1, Treatment I-RF (RF denotes replication with field values) simply manipulates the environment in Treatment I-R by setting

Seller payoff: $\prod_s = p - c(q)$ (2)

Buyer payoff: $\prod_b = v(q) - p$ $p \in [\$5, \$80]$, $q \in [1,5]$,

where c(q) = \$4, \$5, \$8, \$15, and \$50 for q = 1, 2, 3, 4, 5 and v(q) = \$6, \$8, \$15, \$30, and \$80 for q = 1, 2, 3, 4, and 5. These values were chosen to represent the dealer cost (c(q)) to replace a 1990 *Leaf* Frank Thomas card of various quality levels and consumer values (v(q)) for various 1990 *Leaf* Frank Thomas cards. The values are taken from the standard price guide for baseball cards—*Beckett Baseball Cards Monthly*. For each single type of ungraded card, Beckett collects pricing information from about 110 card dealers throughout the country and publishes a "high" and "low" price reflecting current selling ranges for several quality variants. The high price represents the highest reported selling price and the low price represents the lowest price one could expect to find with extensive shopping. Thus, for c(q) values I take the "low" prices from Beckett for 1990 *Leaf* Thomas cards that would grade PSA 6, 7, 8, 9, and 10, and for v(q) I take the "high" prices from Beckett for 1990 *Leaf* Thomas cards that would grade PSA 6, 7, 8, 9, and 10. These price vectors represent roughly a 50-100 percent markup for dealers, which is in the range of what List (2004a) reports in his empirical examination of bid/ask prices for similar sportscards in this market.

Importantly, use of these parameter values provides the necessary tension between the dominant strategy and the joint-profit maximization actions, but now buyers can realize monetary losses, a realistic component of many market settings. Under this design, the Nash purely selfish prediction is p* = \$5, and for sellers to send minimal card quality, q* = 1. These actions result in $\prod_s$ = \$1 and $\prod_b$ = \$1. Note that in this case there could be losses of up to \$74 (buyer sends \$80 and receives the lowest quality Thomas card); as in the other laboratory treatments (Treatments I and II), after these treatments were carried out I had subjects participate in other unrelated experiments that did not involve interaction to ensure that they would not leave with negative cash balances.

Treatment I-RF1 (RF1 denotes replication with field values in a purely one-shot setting) is identical to Treatment I-RF in every manner except that it is not executed over five periods with five different partners; rather it is a one-shot game. Since in the above treatments, by design

subjects should have construed the setting as one-shot, Treatment I-RF and Treatment I-RF1 should yield similar data patterns if (i) subjects interpret Treatment I-RF as several one-shot games and (ii) experience does not unduly influence play. In total, Treatment I yields 77 data points for buyers and 77 data points for sellers.

Moving to row 2 in Table 1, Treatment II adds context to Treatment I-RF1. In this case, rather than buyers and sellers transacting with abstract commodities, Treatment II adds context that closely resembles the subjects' naturally occurring environment. For example, buyers make an offer to a seller to buy one 1990 *Leaf* Frank Thomas baseball card and the buyer requests a certain PSA grade. Similar to Treatment I-RF1, sellers have five PSA grades available (PSA 6, 7, 8, 9 or 10) and subsequently choose the quality of the Frank Thomas baseball card to give the buyer if they accept the buyer's offer.[9] If one ignores the artificiality invoked by the laboratory experimental setting, this particular treatment provides an environment more closely related to the actual decision making processes in the marketplace from which these subjects are drawn. And, this treatment provides a test of whether context matters. Treatment II includes 32 buyers and 32 sellers.

Treatment III moves the exploration out of the laboratory and into the market where these agents actually consummate business: the floor of the sportscard show. Treatments III$20 and III$65 represent the naturally occurring analogues to Treatment II. In these treatments, I have buying agents approach dealers on the floor of a sportscard show and purchase 1990 *Leaf* Frank Thomas baseball cards.[10] Each participant's experience typically followed four steps: (1)

[9] PSA grades 6-10 were chosen because little trading of Thomas cards below PSA 6 is carried out.

[10] As I have noted elsewhere (e.g., List, 2004b, 2004c), with the rise in popularity of collector sportscards and memorabilia over the past two decades, markets that organize buyers and sellers have naturally arisen. Temporal assignment of the physical marketplace is typically done by a professional association or local sportscard dealer, who rents a large space, such as a gymnasium or hotel conference center, and allocates six-foot tables to dealers for a nominal fee. When the market opens, consumers mill around the marketplace, haggling and bargaining with dealers, who have their merchandise prominently displayed on their six-foot table. The duration of a typical sportscard show is a weekend, and subjects enter the market ready to buy, sell, and trade.

consideration of the invitation to participate in an experiment, (2) learning the market rules, (3) actual market participation, and (4) conclusion of the experiment and exit interview.

In Step 1, potential subjects approached the monitor's dealer table and inquired about purchasing late 1980s/early 1990s baseball cards displayed on the table. If the subject was a white male roughly 25 years in age, the monitor asked if he was interested in participating in an experiment that would last about 30 minutes.[11] If the agent agreed to participate, in Step 2 a monitor thoroughly explained the experimental rules. The agent was informed that he would be a "buyer" of 1990 *Leaf* Frank Thomas baseball cards in the experiment. This particular card was chosen due to my experience in evaluating the attributes of the card over the past 15 years (as a dealer and consumer), Thomas' popularity, and the fact that this variant represents his "rookie card"—typically a player's most sought after card. These latter two factors help to explain the extensive interest in the card among broad classes of collectors.

The agent was told that he would approach five different dealers on the floor of a sportscard show to purchase the Thomas card. I was able to pre-select the dealers to be approached before the show by visiting their dealer table and examining whether they had a fair number (more than 5) of Thomas ungraded 1990 *Leaf* cards for sale that were of sufficiently heterogeneous quality. It is common practice for dealers to mill around the show looking at others' goods, and I was merely behaving in accordance with this norm when visiting dealer tables.

Importantly, in the spirit of the literature that suggests contracted negotiations can crowd out reciprocity (see, e.g., Fehr and List, 2004), I was careful to instruct buying agents to avoid haggling, while keeping the transaction as natural as possible.[12] In practice negotiations are typically quite

[11] Given the results in List (2004a), I wished to avoid any confounds associated with statistical discrimination in this marketplace; hence I opted to use "majority" subjects as my buying agents in all treatments. This design choice may well give social preferences their best chance since the data in List (2004a) suggest that these buying agent types receive the best offers from dealers. Note, however, that any agent who desired to participate in an experiment was able to do so since the minority agents were asked to participate in an unrelated pilot experiment.

[12] See also Macaulay (1963), who reports that "detailed negotiated contracts can get in the way of creating good

short or do not occur at all in this market (see List, 2004a, Table II); thus, besides realism this approach gives social preferences their best shot, since buying agents are signaling a fair amount of trust in the dealer when purchasing non-graded sportscards without much detailed negotiations. To ensure that buying agents did not aggressively bargain, their payoffs were not tied to quality or price; rather, they were paid a flat rate of $20 for approaching five dealers. Finally, to maintain consistency with Treatment II, the buying agent offered $20 (or $65) and requested a 1990 *Leaf* Thomas card that would merit a PSA 9 (10) if graded.

In Step 3, the subject approached dealers one at a time. Each interaction lasted less than 3 minutes and resulted in the purchase of a Thomas *Leaf* sportscard. It should be noted that throughout the experiment the sportscard dealers were not aware that an experiment was occurring. This ensured that the process was as natural as possible for the dealers, whose behavior is of primary interest in this field experiment. Step 4 concluded the experiment—after subjects completed a confidential survey, they were paid $20 in private (Appendix B contains the survey).

A few noteworthy design issues should be mentioned before proceeding. First, each dealer was approached twice: once in Treatment III$20 and once in Treatment III$65. The spacing of visits was such to attenuate any suspicion—one example is that dealer *i* was approached by agent *n* on Friday night and by agent *m* on Sunday morning. This aspect of the design provides considerable statistical power, as I can observe within- and between-dealer behavior. And, the ordering of the visits was random—some dealers were approached in the $20 treatment first, others were approached in the $20 treatment second; in practice I observed no ordering effect, so I suppress further discussion of this issue.

---

exchange relationships between business units," and Sitkin and Roth (1993, p. 376), who assert that "legalistic remedies can erode the interpersonal foundations of a relationship they are intended to bolster because they replace reliance on an individual's good will with objective, formal requirements."

Second, unlike audit studies that test for market discrimination, in these treatments I am actually directing the agent to buy the good. In this sense, these are not transactors who obliquely discontinue bargaining if the dealer accepts an offer; these are actual transactions. And, since transactions are typically in cash at sportscard shows, I provided the necessary funds to purchase the cards. Third, note that great care was taken to ensure that the data were gathered from interactions that would naturally occur in the marketplace. Subjects were entering the market to buy goods that were very similar to the good that I had them buying. Fourth, these treatments were carried out at sportscard shows in the same region in the U.S., from October 2002 to July 2003.

Fourth, parameter values in Treatment III were guided by the results in Treatment II and current sportscard market values. Since the average buying agent sent $20 to dealers in Treatment II and requested a PSA 9 Thomas card, Treatment III$20 is the naturally occurring analogue. Treatment III$65 used the same dealers who were visited in Treatment III$20, and was identical in every sense except that in this case buying agents offered $65 for the Thomas card and requested a PSA 10. I chose $65 because it is roughly 33 percent greater than c(10) = $50, matching the relationship of c(9) = $15 and the $20 value chosen in Treatment III$20.

Since quality is difficult to detect in this market for untrained consumers, if social preferences play a role in this case, then the card's grade and the price offer should be positively correlated. Once the buying agents had purchased each of the cards in these treatments, the last step was to have the cards professionally graded. This was completed by having every card graded by a PSA representative.

In total, I observe the behavior of 50 dealers who were each visited by two different agents (one in Treatment III$20 and one in Treatment III$65) — thus I have a sample size of 100 in Treatment III. Similar to nondealers, in every case I was able to obtain important subject-specific information from the dealers, either via a survey they completed during an experiment in which

they later participated or through filling out a survey in exchange for a payment of $1. Appendix B provides a summary of the information that I obtained from dealers.

To explore a level deeper into the underlying structure that organizes behavior in this market, I complete three final treatments making use of natural exogeneity that the market offered during the sample period: while a third-party (PSA) has graded sportscards since 1987, no service existed prior to June of 2003 to grade sporting event tickets and ticket stubs. PSA announced their grading intentions in April 2003, but they did not provide grading criteria until June 2003. As noted earlier, Brown et al. (2004) highlight the attractiveness of such natural variation by arguing that such exogeneity is impossible to find in field data. I believe that these three field experimental treatments offer this useful characteristic.

Treatment IV-NG (denotes no grading available) is identical to Treatment III in that buyers approached dealers on the floor of a sportscard show (from October 2002 to March 2003) with either $10 or $30 to purchase an unused ticket or ticket stub. Given the thinness of the ticket market, it was necessary to use five different ticket types in the purchasing tasks (Cal Ripken's last game at Camden Yards, Cal Ripken's final game of "The Streak," Cal Ripken's "consecutive world-record breaking" game, and two World Series games). I was careful to choose tickets that were in the same price range to increase the likelihood of having the luxury of pooling the data. In total, I observe the behavior of 30 dealers in this treatment and therefore gather 60 data points since each dealer is approached twice.

Treatment IV-AG (denotes after announcement of grading) was completed at sportscard shows after PSA announced they would begin grading ticket stubs (April 2003) but before they released their grading scheme (June 2003). In this treatment, I purchased the same tickets and used the identical protocol as in Treatment IV-NG. As outlined in row 4 column 2 in Table 1, I observe 54 dealer decisions in this treatment.

Completing the experimental design is Treatment IV-G (denotes grading available), which is identical to Treatments IV-NG and IV-AG, but was completed post-June 2003. I observe 36 total dealer decisions in this final treatment. Accordingly, I purchased 150 tickets in Treatment IV; and similarly to Treatment III, I subsequently had every ticket graded by a PSA representative.

Sportscard Grading

Before proceeding to the results summary, it is important to provide the necessary institutional details to motivate the study appropriately. Each year, sportscard companies design and print sets of sportscards depicting players and events from the previous season. Once the print run of a particular set has been completed, the supply of each distinct card in the set is fixed. The value of a particular card depends on its scarcity, the player depicted, and the physical condition of the card—i.e., condition of the edges, corners, surface, and centering of the printing. To track card condition, people often use a 10-point scale. For example, a card with flawless characteristics under microscopic inspection would rate a perfect "10", while defects, including minor wear on the corners, would decrease the card's grade to a "7". The card's overall grade is computed via the aggregation of the various characteristics.

PSA (Professional Sports Authenticators) is the industry leader in grading services, and its parent company became publicly traded in 1999 (Collectors Universe, under NASDAQ ticker symbol CLCT). PSA has graded more than 7 million sportscards since its inception in 1987. Professional grading is voluntary and costs $6-$100 per card, depending on package size and requested turnaround time. Importantly, the fee is independent of the actual grade received. Graded cards are encased in plastic and sealed with a sonic procedure that makes it virtually impossible to open and reseal the case without evidence of tampering.

PSA adopted integer grades from 1 to 10, where a "10" is considered Gem Mint and commands a premium price. A PSA "9" card is considered Mint and is the next most valuable card

type. As witnessed by the c(q) and v(q) vectors used in Treatments I and II, card values are convex in the grade received. Importantly, Jin et al. (2004) provide evidence suggesting that even under PSA's coarse grading system, certification reveals important information to ordinary consumers. Yet they report that dealers gain no information from a card's PSA grade, suggesting that dealers are able to evaluate quality as well as PSA.

Sports tickets and ticket stubs have recently gained enough market acceptance to merit professional grading. Ticket supply, of course, depends on the stadium size of the event and the proportion of fans in attendance that preserved their ticket stubs (or in the case of unused tickets, the number of fans who left their tickets unused). Ticket grading is similar to sportscard grading: an identical 10-point scale is used, and sharpness of corners, centering of ticket, sharp focus, and original gloss are very important. Furthermore, staining, printing imperfections, and print quality of crucial game information are also important in determining ticket quality.

III. Experimental Results

Table 2 provides a summary of the raw data. The table can be read as follows: Treatment I-R in row 1, column 1, denotes that the average price in this treatment was $28.40, average quality was 3.5, and average requested quality was 6.1. Note that in Table 2, for comparability reasons, I have scaled Treatment I-R data to range from 1-10, and PSA 6, 7, 8, 9, and 10 are denoted as quality levels 1, 2, 3, 4, and 5.[13] A first result relates to the comparison between the behavior of this subject pool and students. As Fehr and List (2004) note, a typical criticism levied against experimental results concerns the fact that most economics experiments are conducted with students. This may be problematic for several reasons. For example, due to selection effects, those

[13] Average individual payoffs (ranges of individual payoffs) in the laboratory treatments are as follows: Treatment I-R: buyers, $14.90 ($6.5 to $24), sellers, $18.60 ($5 to $34); Treatment I-RF: buyers, $2.40 (-$59 to $25), sellers, $8.00 ($1 to $61); Treatment I-RF1: buyers, $0.22 (-$25 to $25), sellers, $9.81 ($1 to $35); Treatment II: buyers, -$0.09 (-$67 to $25), sellers, $8.44 ($1 to $70).

who do not behave like students may have selected into roles and be overrepresented in certain parts of the economy (e.g., sellers in the marketplace).[14] The first result addresses this issue.

Result 1: *Behavior of sportscard enthusiasts in laboratory games is consistent with the gift exchange literature using student subjects, and the results extend well to one-shot environments.*

Evidence for Result 1 is contained in the raw statistics in row 1 of Table 2, which are similar to the raw data gathered in laboratory experiments with student subjects. Figure 1 complements Table 2 by mapping the relationship between product quality and prices for Treatment I-R. Figures 2 and 3 provide similar insights using data from Treatments I-RF and I-RF1. Overall, the trajectory of the data clearly shows that product quality and prices are positively related. In addition, when I examine the temporal aspect of the data there is little variation over time, consistent with previous studies on gift exchange (for an exception, see Charness et al., 2004).

To explore these differences further, I estimate Tobit and Tobit random effects regression models using the data from Treatment I. The dependent variable in the regressions is the quality of the good, which is regressed on the price transfer and controls for time and dealer-specific effects:

$$q_{it} = \beta p_{it} + \omega_{it}\,. \qquad (3)$$

In equation (3), q_{it} represents the product quality that dealer i sent to the buyer in period t; p_{it} denotes the buying agent's offer price to dealer i in period t; and ω_{it} includes a constant and a time trend in the Tobit model. This specification is augmented by inclusion of dealer-specific random effects in the Tobit random effects regression model.[15]

Regression results presented in columns 1-3 of Table 3 provide evidence that dealers reward buyers for paying higher prices. In each of the three treatments the marginal effect of price is

[14] The general notion of examining whether natural players are different from students is gaining popularity in the economics literature. For example, Cooper et al. (1999) examine the ratchet effect with middle and upper level Chinese managers and Camerer et al. (2003; 2004) report data from a CEO subsample in a beauty contest game.

[15] In Treatment III and IV data, buyer-specific effects were found to be insignificant, which stands to reason since the agents were homogeneous and followed a standard buying procedure.

positive and statistically significant at the $p < .10$ level using a two-sided alternative. This result, which is consistent with the received gift exchange literature, is in line with the conjecture that positive reciprocity supports cooperative play and mutually beneficial exchange. When applicable, I also present an estimate of θ in Table 3. θ is equal to $\partial v(q)/\partial P$ and provides a natural benchmark of gift exchange expressed in monetary units. In the case of Treatments I-RF and I-RF1, θ estimates are both significantly different from zero, suggesting that gift exchange occurs at the margin. In terms of economic significance, a θ estimate of 1.3 in Treatment I-RF1 suggests that a \$1 increase in P leads to a \$1.30 increase in reciprocated gift (v(q)).

While these results provide a robustness check of the data gathered in the laboratory with student subjects and represent good news in that the major laboratory results seem to spill over to different subject pools who are commonly engaged in similar exercises in their everyday lives, one can push the comparability notion a bit harder by adding field context to the laboratory environment. Upon doing so, I find

Result 2: *Adding natural context to the experimental instructions does not influence behavior.*

Evidence for this result can be found in Table 2, which shows that average prices and quality levels are only slightly lower than what was found in Treatment I-RF1 (the comparable context-free treatment). Slight behavioral differences are also revealed when comparing Figures 3 and 4, which show i) that the positive relationship remains in the contextual data, but that there is a slightly greater mass at the sub-game perfect equilibrium prediction: 13 of 32 (41%) observations in Treatment II versus 9 of 27 (33%) observations in Treatment I-RF1, and ii) that there is a greater number of price (quality) realizations at \$25 (3) and below in Treatment II. While directionally these differences all point to contextual effects, it is important to note that none of these treatment differences are statistically significant at conventional levels using a test of proportions.

To compare gift exchange on the margin across these two treatments, I return to equation (3) and estimate a Tobit model. For Treatment II data, the marginal price effect is positive and statistically significant at conventional levels—see column 4 in Table 3. It is interesting to note that the marginal effect estimate (0.06) is slightly lower than the marginal effect estimate in Treatment I-RF1 (0.10), and θ is considerably lower: \$0.77 versus \$1.3. Upon pooling these data and estimating equation (3), however, a likelihood ratio test suggests that the homogeneity null should not be rejected, suggesting that behavioral differences do not exist across Treatments I-RF1 and II. Accordingly, the overall pattern of results suggests that gift exchange is alive and well, even when context is included in the experimental instructions.

Results 1 and 2 provide a nice validity check of the extant gift exchange literature. A necessary next step in this line of research is to explore behavior in naturally occurring environments where the controls of the experiment are relaxed. This constitutes one goal of Treatments III and IV, which yield a first insight:

Result 3: *When third-party verification is available, behavior in naturally occurring transactions is consonant with the notion of gift exchange.*

Tables 2 and 3 provide evidence for Result 3. Row 3 in Table 2 shows that the positive relationship between price and product quality is evident in the aggregate data: whereas the average quality was 2.1 (PSA 7.1) in Treatment III\$20, it was 3.2 (PSA 8.2) in Treatment III\$65. In addition, data from Treatments IV-AG and IV-G in row 4 of Table 3 support the positive relationship found in the sportscard data.[16]

Regression results in Table 3 yield similar insights: estimates in column 5 of Table 3 provide evidence that product quality and price are positively correlated in Treatment III, as the marginal effect estimate of 0.02 is positive and significant at conventional levels—this estimate

[16] Using a Wilcoxon signed rank test for matched pairs, I find that all differences are statistically significant at the $p < .05$ level except for Treatment IV-NG data; this result is discussed more fully below.

suggests that card quality increases by roughly 1 grade when the buyer offers \$65 rather than \$20; in this case, θ is equal to \$0.21.[17] A similar result holds in the Treatment IV-AG and IV-G data presented in columns 7 and 8 of Table 3, although the marginal price effect is not statistically significant in the Treatment IV-G data at conventional levels. Upon pooling the Treatment IV-AG and IV-G data (a likelihood ratio test indicates pooling is appropriate—LLR test: $\chi^2 = 5.8$), however, the marginal price effect, contained in the rightmost column of Table 3, is statistically significant. Interestingly, across all three specifications the marginal price effect estimate is 0.02, and θ is approximately \$0.20.[18]

Considering that this data pattern is observationally equivalent to predictions from a purely reputational model that includes no social preferences, one can dig a level deeper into these data by recognizing that some of the dealers in the sample may have had an economic reason to uphold their reputations, whereas others may not have had similar incentives. A next result follows:

Result 4: *When third-party verification is possible, local dealer behavior in naturally occurring transactions is consonant with extant empirical insights concerning social preferences, whereas non-local dealers' behavior is in line with self-interest theory.*

Table 4 and Figures 5 and 6 provide evidence for this result. In splitting the dealer types, a dealer is labeled as a "non-local" if he or she is unlikely to be concerned with reputation effects—for example, rarely attends sportscard shows in the area (fewer than three times in a typical year), does not plan to attend more frequently than this in the future, does not own a sportscard shop, and does not have an Internet sportscard business. All other dealers are labeled as "locals"—in practice, these are primarily dealers who frequent the area often. This information was obtained from the survey summarized in Appendix B.

[17] In addition to the Tobit random effects estimation strategy, which is heavily utilized in the literature, since there is a natural ordering in the data and there are only 5 cells (i.e., PSA 6-10), I supplement these results by using a panel data ordered probit model, as described in Appendix C. Empirical estimates from the panel data ordered probit model are suppressed because they always coincide with insights gained from equation (3).

[18] When computing θ in the ticket specifications, v(q) is equivalent to ½ the value of v(q) in the sportscard data.

The raw data displayed in Figures 5 and 6 provide initial support for Result 4. When dealing with local dealers, higher price offers yield superior quality in Treatments III, IV-AG, and IV-G, as illustrated in Figure 5. Alternatively, while delivered quality is positively related to price across these three treatments among non-local dealers (see Figure 6), the differences are tiny and never statistically significant using a Wilcoxon signed rank test for matched pairs.

Table 4 provides regression results to support Result 4. Columns 1 and 2 split the Treatment III data into two subsamples: III_L (local dealer data) and III_N (non-local dealer data). In the former subsample, the marginal price effect is positive and statistically significant at conventional levels. In terms of economic significance, the coefficient estimate in column 1 of 0.03 results in an estimated marginal effect of roughly 1.5 grades: that is, in the \$65 treatment local dealers provided a quality that was 1.5 grades above the quality level they provided in the \$20 treatment. Measured at the sample means, this 1.5 quality increment yields the buyer a PSA rated 8.6 card rather than a PSA rated 7.1 card. Using the v(q) values discussed earlier, this quality increase maps into an increase in market value of roughly \$20, much less than the extra \$45 spent to obtain the card. A θ estimate of \$0.31 complements this finding.

Alternatively, for non-local dealers gift exchange is not evident in Treatment III (see column 2 of Table 4), as the marginal price effect is not statistically significant at conventional levels. Regression results for Treatments IV-AG and IV-G provide further support for Result 4: in both cases the marginal price effect in the local dealer data is positive and significant at conventional levels (columns 5 and 7 of Table 4), whereas there is no such effect found in the non-local dealer data (columns 4 and 6 of Table 4). For both the Treatment IV-AG_L and IV-G_L data, the marginal effect estimate is 0.04, and θ is \$0.32 and \$0.42, though neither θ estimate is statistically significant at conventional levels. Upon pooling the Treatment IV-AG_L and IV-G_L data (LLR test: $\chi^2 = 1.4$), θ equals \$0.35 and is significant at the $p < .05$ level (rightmost column of Table 4). Treating non-

local dealer data similarly by pooling Treatment IV-AG_N and IV-G_N provides little new information: gift exchange is not evident among non-local dealers.

A natural question that arises concerns whether the local dealer behavior is driven primarily by reputation effects or social preferences—given the identification problem, from the above results alone one cannot determine the extent to which reputation effects and social preferences are influencing the outcomes. One nice characteristic of the current experimental design is that I can examine behavior in markets that are void of third-party verification to explore this issue. In such cases, in economic terms the situation faced by the local and non-local dealers is identical. Treatment IV-NG provides a first result:

Result 5: *When third-party verification is not available, supply side behavior in naturally occurring transactions is consonant with purely selfish money-maximizing theory.*

Evidence for this result can be seen in Tables 2-4 as well as Figures 5 and 6. Table 2 shows that there is very little quality difference between the \$10 and \$30 offers in Treatment IV-NG. Indeed, this quality difference is not statistically significant using a Wilcoxon signed rank test for matched pairs. This result is highlighted in Figures 5 and 6, where both local and non-local dealers do not provide different quality levels across offers of \$10 and \$30 in Treatment IV-NG. Empirical results displayed in Tables 3 and 4 support the raw data patterns, as the marginal price effect is insignificant in the aggregate data (column 6 in Table 3) and in both regressions that split the data by dealer type (columns 3 and 4 in Table 4).

This finding leads to the tentative conclusion that reputation effects rather than social preferences are responsible for driving a large part of the price/quality tendencies observed in the naturally occurring data. While certainly there is some evidence in favor of social preferences in this market, as directionally it is evident in various places in the non-local dealer data and in the local dealer Treatment IV-NG data, it seems to be of second-order importance in real market transactions.

Clearly, understanding these types of market transactions is important since they replicate the one-shot transactions of laboratory experiments and are prevalent in many naturally occurring settings, but oftentimes long-term relationships can form in markets. This is the case in certain labor markets and in some product and service markets as well. Given that I also gathered data on the nature of previous interactions (see Appendix B), it is possible to determine whether outcomes in transactions that are part of a long-term relationship provide evidence consistent with social preferences. In doing so, an interesting result follows:

Result 6: *For transactions within long-term relationships there is evidence consistent with social preferences.*

Primary evidence for this result can be obtained from the Treatment IV data. First, it is important to note that regression models that pool the Treatment III and IV data and include an indicator variable for whether the buyer and seller had previous interactions yield estimates in line with Result 6: in those cases where the dealer and buyer had previous interactions, delivered quality is considerably higher, *ceteris paribus*. Of course, this evidence alone is not strong because reputation effects and social preferences are both elements in these transactions.

To examine reputation effects in isolation, I estimate equation (3) using Treatment IV-NG_L data, but augment the specification in column 3 of Table 4 by including an interaction term: price*previous interaction, where previous interaction equals 1 if the buyer and dealer have had five or more interactions in the previous 12 months or have had two or more interactions annually over the past 3+ years, and equals 0 otherwise. I observe 12 such pairs in the Treatment IV-NG_L data and label these pairs "long-term" relationships. Estimation results, suppressed for parsimony, yield a zero coefficient estimate on price and a positive coefficient estimate on the interaction term that is significant at the $p < .05$ level. In terms of economic significance, the increase in price from \$10 to \$30 in long-term interactions yields an estimated increase in product quality of 0.40 grades. If one assumes that reputation effects in such transactions are nil, this estimate provides a measure of

social preferences within long-term relationships. To put this estimate into perspective, one can compare this marginal price effect with the estimated quality increase in the IV-AG_L and IV-G_L data among long-term interactions. Using an identical identification strategy, I find that in these cases the marginal price effect is equivalent to 1.26 PSA grades. Thus, considering the empirical results for the non-local dealer data, a rough estimate is that about one-third of the 1.26 quality grade increment is due to social preference effects and the other two-thirds is most likely due to reputation effects.

The above empirical estimates provide measures of social preference effects and reputation effects, but it is also important to recognize the degree of mendacious claims in the marketplace. If dealers do not have the necessary inventory to fulfill the quality request (for example, due to my misjudgment of quality during my walk-by or due to sales during the show) but provide quality disclaimers, then it is important to recognize and explore this aspect of behavior. In this sense, a first result follows:

Result 7: *When third-party verification is possible, local dealers provide fewer claims of quality than non-local dealers, and conditional on claiming quality, shirk less frequently.*

Table 5 summarizes dealer behavior across the various treatments. The first part of Result 7 can be obtained by computing the percentage of local and non-local dealers who claim quality in Treatments III, IV-AG, and IV-G. The second part of Result 7 follows from a comparison of the quality claimed and the quality actually delivered. Before discussing the evidence for Result 7, it is important to point out that in some cases dealers provide quality ranges – for example, "this card would grade at PSA 8 or 9." In these cases I use the mid-point of the range (e.g., 8.5). A few other dealers were agnostic about the grading system—I label these types as not claiming quality (similar results are obtained if I simply delete these observations). And, in some instances the dealer stated "this one is top quality" or "this is a gem" when describing the good. I label these dealers as not

claiming quality, but should note that if I take the literal word of the dealer and pair these statements with the appropriate PSA grade the fundamental results do not change.

Upon pooling the Treatment III, IV-AG, and IV-G data in Table 5, I find that 94 of 190 (49%) dealer observations involve product quality claims. Split by dealer type, 38 of 120 (32%) local dealer observations involve product quality claims, whereas 56 of 70 (80%) non-local dealer observations involve product quality claims. These proportions are statistically different at the $p < .05$ level using a test of proportions.[19] Of those dealers who make quality claims, local dealers deliver the promised quality (or above) in 18 of 38 cases (47%), whereas non-locals deliver the promised quality (or above) in only 5 of 57 (9%) cases. Using a test of proportions, I find that these percentages are significantly different at the $p < .05$ level.

Similar to the spirit of the inquiry into Result 4, one can question whether the increased quality promises and deliveries from local dealers are due purely to reputational concerns or have an element of social preferences. Examining Treatment IV data lends insights into this issue and leads to the next result:

Result 8: *When third-party verification is not possible, local and non-local dealers make similar claims of quality, and conditional on claiming quality, shirk to the same extent.*

As Table 5 reveals, in Treatment IV-NG local dealers make quality claims in 22 of 36 (61%) cases, whereas non-local dealers make quality claims in 14 of 24 (58%) cases. This difference is not statistically significant at conventional levels. Likewise, conditional on claiming quality, local dealers in Treatment IV-AG shirk in 18 of 22 cases—i.e., in 82% of transactions local dealers provide lower quality than promised—whereas 71% (10 of 14) of non-local dealer observations

[19] As Table 5 illustrates, results are similar if I analyze the treatments separately. For example, in Treatment III, I find that 26 of 30 non-local dealer observations have quality claims, whereas only 27 of 70 local dealer observations have quality claims. These proportions are different at conventional significance levels. Note that these observations are non-independent within a treatment type—in some cases dealers make 2 quality claims (once in the low price treatment and once in the high price treatment). In these cases, I average the quality claims to ensure independence in the statistical tests.

should be considered shirking. Again, this result is not statistically significant at conventional levels.

Interestingly, while quality claims and shirking rates are not considerably different for non-local dealers across Treatments III and IV, they are considerable different for local dealers. Among local dealers, *more* claims of quality and *higher* shirking rates are evident when third-party verification is not possible. This insight can be obtained via comparison of the local dealer data in Treatment IV-NG with the local dealer data in the other three treatments (row 2, column 2, versus row 2, columns 1, 3, and 4).

Overall, these results complement Results 1-7 yet it is important to consider outcomes in long-term relationships considering the insights gained from Result 6. Doing so yields the following result:

Result 9: *When third-party verification is not possible, local dealers within long-term relationships make more claims of quality, and conditional on claiming quality, shirk less often than when they are outside of long-term relationships.*

Evidence for this result can be obtained from the Treatment IV data. First, in Treatment IV-NG, local dealers make product quality claims in 75% (9 of 12) of deals within long-term relationships, much higher than the rate of 54% (13 of 24) of claims that local dealers make outside of long-term relationships. In terms of shirking rates, the insights gained from the Treatment IV-NG data paint a picture similar to Result 6: 56% (5 of 9) of local dealer observations should be considered shirking in long-term relationships, whereas 100% (13 of 13) of the local dealer observations should be considered shirking when they are not part of a long-term relationship. This finding complements Result 6 and suggests that in long-term relationships social preferences influence actual transactions in this marketplace.[20]

[20] When I examine local dealer data in Treatments III, IV-AG, and IV-G, I find that in long-term relationships they shirk considerably less often (roughly 14% of observations) than local dealers shirk outside of long-term relationships (roughly 56% of observations).

In light of these results, one might expect that rational buying agents do in fact refrain from purchasing ungraded products from strangers. This hypothesis can be examined by returning to the laboratory data and more fully exploring the nature of price offers across experienced and inexperienced buying agents. A general insight follows:

Result 10: *Experienced buying agents exercise caution when product quality is uncertain, leading buyer-side behavior among the experienced agents to be in line with purely selfish money-maximizing theory.*

Evidence for this result is obtained by regressing offered price on a vector of buyer-specific variables including individual experience levels. Only summarized here for brevity, in each of the empirical models, market experience, and more specifically experience with professionally graded cards, leads to lower levels of price transfers at conventional significance levels. This result is important in the sense that it suggests that buying agents might learn to avoid deals that involve lower quality products, and suggests that, with proper information dissemination, in long-run equilibrium few ungraded products will exchange hands among strangers.

IV. Conclusions

This study provides a framework for measuring social preferences and reputation effects using a series of laboratory and field experiments in an actual marketplace. Empirical results suggest that reputational effects are quite important in this particular market and that social preferences do not considerably influence outcomes, except in those interactions within a long-term relationship—in those cases where the dealer and buyer had previous interactions, delivered quality is considerably higher. This result is consistent with Brown et al.'s (2004) laboratory results. Empirical results also suggest that third-party enforcement of contracts is important, even among agents who seemingly have social preferences. This result follows from the increased level of delivered product quality when third-party enforcement was available.

One thought for future work is to examine whether an explicit consumer threat to return the good if it does not grade according to the quality claimed affects shirking rates—in those cases where social preferences are found to be prevalent, it may backfire by inducing less trustworthy behavior. Accordingly, incentives that explicitly threaten to penalize shirking may involve hidden costs. In recent years, economists have focused attention on similar phenomena (e.g., Benabou and Tirole 2002). This discussion will be reserved for another occasion.

References

Akerlof, George A. (1982), "Labor Contracts as Partial Gift Exchange," *Quarterly Journal of Economics*, 97, 543-569.

Akerlof, George A. and Janet L. Yellen (1988), "Fairness and Unemployment," *American Economic Review, Papers and Proceedings* 83(2), 44-49.

Akerlof, George A. and Janet L. Yellen (1990), "The Fair Wage-Effort Hypothesis and Unemployment," *Quarterly Journal of Economics*, 105, 255-283.

Andreoni, James and John Miller (2002), "Giving according to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica,* 70, 737-753.

Benabou, Roland, and Tirole, Jean (2002). "Self-Confidence and Personal Motivation," *Quarterly Journal of Economics*, 117, 871-915.

Berg, Joyce, John W. Dickhaut, and Kevin A. McCabe (1995), "Trust, Reciprocity, and Social History," *Games and Economic Behavior* 10, 122-142.

Bewley, Truman F., (1995). A Depressed Labor Market as Explained by Participants. *American Economic Review, Papers and Proceedings* 85(2), 250-254.

Blinder, Alan S. and Don H. Choi (1990), "A Shred of Evidence on Theories of Wage Stickiness," *Quarterly Journal of Economics*, 105, 1003-16.

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

Brown, Martin, Armin Falk, and Ernst Fehr (2004), "Relational Contracts and the Nature of Market Interactions," forthcoming *Econometrica*.

Camerer, Colin (2002). Behavioral Game Theory. Princeton, Princeton University Press.

Camerer, Colin and Keith Weigelt (1988). Experimental Tests of a Sequential Equilibrium Reputation Model, *Econometrica*, 56, 1-36.

Camerer, Colin, Teck-Hua Ho, and Juin-Kuan Chong (2003). Models of Thinking, Learning, and Teaching in Games. *American Economic Review, Papers and Proceedings* 93(2), 192-195.

Camerer, C., Teck-Hua Ho, and Juin-Kuan Chong (2004). A Cognitive Hierarchy Model of Games, forthcoming, *Quarterly Journal of Economics*.

Charness, Gary (1996), "Attribution and Reciprocity in an Experimental Labor Market," mimeo.

Charness, Gary and Matthew Rabin (2002), "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117, 817-870.

Charness Gary, Guillaume R. Frechette, and John H. Kagel (2004), How Robust is Laboratory Gift Exchange?" *Experimental Economics*, forthcoming.

Chiappori, P. A. and Bernard Salanie (2003). "Testing Contract Theory: A Survey of Some Recent Work", in: Mathias Dewatripont, Lars Peter Hansen, Stephen J. Turnovski, *Advances in Economic Theory, Eighth World Congress of the Econometric Society*, Cambridge: Cambridge University Press.

Cooper, David J., John H. Kagel, W. Lo, and Qing Liang Gu (1999). "Gaming against Managers in Incentive Systems: Experiments with Chinese Students and Chinese Managers", *American Economic Review*, 89, 781-804.

Cox, James C. (2004). "How to Identify Trust and Reciprocity," *Games and Economic Behavior*, 46, 260-281.

Dufwenberg, Martin and Georg Kirchsteiger (1999). "A Theory of Sequential Reciprocity," Discussion Paper, CentER, Tilburg University.

Falk, Armin and Urs Fischbacher (1999). A Theory of Reciprocity. Institute for Empirical Research in Economics, Working Paper No. 6, University of Zürich.

Fehr, Ernst and Armin Falk (1999), Wage Rigidity in a Competitive Incomplete Contract Market, *Journal of Political Economy*, 107, 106-134.
Fehr, Ernst and Simon Gächter (2000), Fairness and Retaliation: The Economics of Reciprocity, *Journal of Economic Perspectives*, 14(3), 159-181.
Fehr, Ernst and Simon Gächter, and George Kirchsteiger (1997). Reciprocity as a Contract Enforcement Device, *Econometrica*, 65 (4), 833-860.
Fehr, Ernst, George Kirchsteiger, and Arno Riedl (1993). "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics* 108, 437-460.
Fehr, Ernst, and John A. List, (2004). "The Hidden Costs and Returns of Incentives – Trust and Trustworthiness among CEOs," forthcoming, *Journal of the European Economic Association*.
Fehr, Ernst, and Klaus Schmidt, (1999). A Theory of Fairness, Competition, and Cooperation, *Quarterly Journal of Economics* 114, 817-868.
Gächter, Simon, and Armin Falk, (2002). "Reputation and Reciprocity: Consequences for the Labour Relation, *Scandinavian Journal of Economics*, 104(1), 1-27.
Hannan, R. Lynn, John H. Kagel, and Donald V. Moser (2004). "Partial Gift Exchange in an Experimental Labor Market: Impact of Subject Population Differences, Productivity Differences, and Effort Requests on Behavior," *Journal of Labor Economics* 20(4), 923-951.
Holt, Charles A. and Roger Sherman (1990) "Advertising and Product Quality in Posted-Offer Experiments," *Economic Inquiry*, 28, 39-56.
Jin, Ginger, Andrew Kato, and John A. List (2004). "Evolution of Professional Certification Markets: Evidence from Field Experiments," working paper, University of Maryland.
Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler (1986). "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, 76, 728-741.
Katz, Lawrence. F. (1986). "Efficiency Wage Theories: A Partial Evaluation," in *NBER Macroeconomics Annual*, ed. S. Fischer, Cambridge, MA: MIT Press.
List, John A. (2004a). "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *Quarterly Journal of Economics*, 108, 45-90.
List, John A. (2004b). "Neoclassical Theory Versus Prospect Theory: Evidence from the Field," *Econometrica*, 72(2), 615-625.
List, John A. (2004c). "Testing Neoclassical Competitive Theory in Multi-Lateral Decentralized Markets," forthcoming, *Journal of Political Economy*.
Lynch, Michael, Ross M. Miller, Charles R. Plott, and Russell Porter (1991). "Product Quality, Informational Efficiency, and Regulations in Experimental Markets," in R.M. Isaac (ed.), *Research in Experimental Economics* 4, Greenwich, CT: JAI Press.
Macaulay, Stewart (1963). "Non-Contractual Relations in Business: A Preliminary Study," *American Sociological Review*, 28, 55-70.
Miller, Ross M. and Charles R. Plott (1985). "Product Quality Signaling in Experimental Markets," *Econometrica* 53, 837-872.
Prendergast, Canice (1999). "The Provision of Incentives in Firms," *Journal of Economic Literature*, 37, 7-63.
Rabin, Matthew, (1993). "Incorporating Fairness into Game Theory and Economics," *American Economic Review* 83, 1281-1302.
Sitkin, S. and Nancy L. Roth (1993). "Explaining the limited Effectiveness of Legalistic Remedies for Trust/Distrust," *Organization Science*, 4, 367-394.
Sobel, Joel (2002). "Social Preferences and Reciprocity," mimeo, University of California, San Diego.
Stigler, George (1981). "Economics or Ethics?" in S. McMurrin, ed. *Tanner Lectures on Human Values*. Cambridge: Cambridge University Press.

Appendix A. Summary Experimental Instructions: Treatment I-R

Welcome to an economics experiment. The instructions are simple, and if you read them carefully and make appropriate decisions, you can earn a considerable amount of money. All of the profits you make in this experiment and other subsequent experiments carried out today will be summed and paid to you in cash in private.

This experiment consists of two stages:

1. 5 non-dealers are buyers, 5 dealers are sellers. You will notice the room divider that separates buyers from sellers—please do not attempt to see who is on the other side of the divider. In the first stage, a buyer will make an offer to a seller to buy one unit of a fictitious good. The buyer will also request a certain product quality.
2. In the second stage, the seller decides whether to accept the price offer and decides on the product quality.

Some important pieces of information:

A. This same decision problem will take place five times (or five periods). Note, however, that buyers will be paired with a different seller each time, thus you are never paired with the same person twice. Earnings will be computed by randomly selecting one of the five periods for payment—the chosen period will be carried out for cash.

B. Sellers cannot make counteroffers—they merely decide whether or not to accept the offer and the product quality.

C. Profit of a buyer is the difference between \$80 and the price at which he has bought the good. This difference is then multiplied by the quality of the good chosen by the seller. Thus, the formula to compute buyer profits is:

($80 – price paid)*quality

D. If buyers decide to make an offer to buy the good from the seller, they must choose a whole number between (or including) \$5 and \$80.

E. The buyer writes down this price offer and a requested quality on the decision sheet provided. Do not announce your decision publicly.

F. After writing these choices on the sheet, a monitor will take the sheet to a seller, who views the choice and decides whether or not to accept the offer. If he accepts the offer, he then determines product quality, which must be between (or including) 0.1 and 1. Sellers are not required to provide the quality that the buyer requested. Sellers should write down their choices on the sheet (whether to accept or not, and product quality), after which the sheets are returned to buyers to compute profits.

G. Seller's profits are given by:

Price paid by the buyer – cost of providing the good

If sellers do not accept the offer, both the buyer and seller receive \$0 for that period. Seller's costs depend on their choice of product quality, as follows:

Quality	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Cost	0	\$1	\$2	\$4	\$6	\$8	\$10	\$12	\$15	\$18

The table shows that the highest quality good (quality = 1) costs the dealer \$18 to provide. A quality of 0.1 costs the dealer \$0. For buyers, if the chosen quality is 1.0, then their profits are simply \$80 – price paid. Otherwise, this difference (\$80 – price paid) is multiplied by a fraction less than 1.

Are there any questions? Let's now go over a few practice problems to ensure everyone understands the rules and how to compute payoffs.

Appendix B. Confidential Survey Summary

These questions will be used for statistical purposes only. THIS INFORMATION WILL BE KEPT STRICTLY CONFIDENTIAL AND WILL BE DESTROYED UPON COMPLETION OF THE STUDY.

1. How long have you been active in the sportscard and memorabilia market? _____yrs

1a. How often do you have your sportscards professionally graded?

Always Sometimes Rarely Never

2. Are you a sportscard or sports memorabilia professional dealer?_____

2a. If yes, in a typical month, how often do you visit this area as a sportscard dealer?_____

2b. As a dealer, how often do you plan to set up in this area in the coming months? _____

2c. Do you have an Internet sportscard/memorabilia business? _____

2d. Do you sell on *eBay* or on a different Internet site using your dealer name?_____

2e. If yes, how often do you sell?_____

2f. Do you own a sportscard shop?_____

3. Gender: 1) Male 2) Female

4. Age _____ Date of Birth _____

Additional questions for Treatment III [Treatment IV]

5. Have you had previous interactions with that dealer? _____If yes, how many?_____ Over how many years?_____.

6. Did the dealer provide any "guarantee" about the PSA grade of the card [ticket]? For example, did the dealer state that "this card [ticket] would grade at PSA 9 [if such services were available]"? Please comment. _____

_____

Appendix C. Alternative Estimation Strategy: Panel Data Ordered Probit Model

I begin by coding dealer behavior for quality: "top quality—PSA 10" "nice quality—PSA 9," etc. Considering these classifications as a ranking of "the propensity to provide quality," I build a model around a latent regression of the form:

$$Q^* = Z'\beta + \varepsilon, \qquad (4)$$

where Q^* is the unobserved vector of "propensity to provide quality," Z is a vector of dealer-specific variables that also includes p, β is the estimated response coefficient vector, and ε_i is the well-behaved random error component. Although I do not directly observe Q^*, I do observe an approximation of Q^*:

$$\begin{aligned} Q = \quad & 6 \text{ if } Q^* \le 0 \qquad (5) \\ & 7 \text{ if } 0 < Q^* \le \phi_1 \\ & 8 \text{ if } \phi_1 < Q^* \le \phi_2 \\ & 9 \text{ if } \phi_3 < Q^* \le \phi_4 \\ & 10 \text{ if } \phi_4 < Q^* \le \phi_5, \end{aligned}$$

where ϕ_i are unknown parameters that are estimated jointly with β. As such, when estimating this model one obtains threshold levels of "propensity to provide quality" by measuring how variables in vector Z affect ranked responses, Q^*.

A few aspects of this particular estimation procedure merit further consideration. First, since the ϕ's are free parameters, there is no significance to the unit distance between the set of observed values of Q, thus avoiding symmetric treatment of one-unit changes in the dependent variable. Second, estimates of the marginal effects in the ordered probability model are quite involved because there is no meaningful conditional mean function. I therefore compute the effects of changes in the covariates on the γ probabilities: $\partial \text{Prob}[\text{cell } \gamma]/\partial Q = [f(\phi_{j-1} - Z'\beta) - f(\phi_j - Z'\beta)]*\beta$ -- where $f(\bullet)$ is the standard normal density, and other variables are defined above. By definition, these effects must sum to zero since the probabilities sum to one.

Empirical estimates from these models are available upon request.

Table 1. Experimental Design

Treatment I	**Treatment I-R** *Replicate lab studies* *n* = 25	**Treatment I-RF** *Extend to field values* *n* = 25	**Treatment I-RF1** *Extend to one-shot environment* *n* = 27
Treatment II	**Treatment II** *Adds context to* I-RF1 *n* = 32		
Treatment III	**Treatment III$20** *Naturally occurring sportscards* *n* = 50	**Treatment III$65** *Naturally occurring sportscards* *n* = 50	
Treatment IV	**Treatment IV-NG** *Naturally occurring tickets before grading was available* *n* = 60	**Treatment IV-AG** *Naturally occurring tickets post-grading announcement* *n* = 54	**Treatment IV-G** *Naturally occurring tickets when grading service is available* *n* = 36

Notes: Each cell represents one (or two, in the case of Treatment IV) unique treatment. For example, Treatment I-R in row 1, column 1, denotes that 25 dealer and 25 nondealer observations were gathered to replicate the laboratory gift exchange studies in the literature.

Table 2. Results Summary

Treatment I	**Treatment I-R** p = 28.4(16.1) q = 3.5(2.0) q_r = 6.1(2.1)	**Treatment I-RF** p = 22.6(20.7) q = 2.3(1.4) q_r = 4.1(0.9)	**Treatment I-RF1** p = 24.8(22.1) q = 2.5(1.7) q_r = 4.0(1.3)
Treatment II	**Treatment II** p = 19.5(19.6) q = 2.3(1.5) q_r = 4.2(1.1)		
Treatment III	**Treatment III$20** p = $20 q = 2.1(0.9) q_r = 4	**Treatment III$65** p = $65 q = 3.2(1.0) q_r = 5	
Treatment IV	**Treatment IV-NG** p = $10 p = $30 q = 2.7(0.6) q = 2.7(0.7) q_r = 4 q_r = 5	**Treatment IV-AG** p = $10 p = $30 q = 2.9(0.6) q = 3.4(0.8) q_r = 4 q_r = 5	**Treatment IV-G** p = $10 p = $30 q = 3.1(0.8) q = 3.6(1.1) q_r = 4 q_r = 5

Notes: Each cell represents summary statistics from one (or two in the case of Treatment IV) unique treatment. For example, Treatment I-R in row 1, column 1, denotes that the average price in this treatment was $28.40, average quality was 3.5, and average requested quality was 6.1. Treatment I-R data are scaled to range from 1-10, and PSA 6, 7, 8, 9, and 10 are denoted as quality levels 1, 2, 3, 4, and 5 in the table. Standard deviations are in parentheses beside means.

Table 3: Marginal Effects Estimates for the Sellers' Quality[a,b]

	Treatment Type								
Variable	**I-R**	**I-RF**	**I-RF1**	**II**	**III**	**IV-NG**	**IV-AG**	**IV-G**	**IV-P**
Price	0.05*	0.05^	0.10^	0.06^	0.02^	-0.001	0.02^	0.02	0.02^
	(1.8)	(3.3)	(5.0)	(4.2)	(6.6)	(0.01)	(2.1)	(1.1)	(2.6)
Constant	0.6	-0.4	-0.8	-0.6	0.6^	1.7^	1.6^	1.8^	1.7^
	(0.7)	(0.7)	(1.7)	(1.7)	(3.1)	(8.0)	(5.8)	(3.3)	(7.3)
θ	---	$0.72^	$1.3^	$0.77^	$0.21^	$0.01	$0.17	$0.23	$0.21^
		(3.6)	(5.5)	(4.2)	(5.0)	(0.3)	(1.1)	(1.1)	(2.3)
Person Random Effects	YES	YES	NO	NO	YES	YES	YES	YES	YES
N	25	25	27	32	100	60	54	36	90

[a]Dependent variable is the sellers' product quality given to the buyer. IV-P pools IV-AG and IV-G data. θ is the monetary gift exchange estimate, computed as $\partial v(q)/\partial P$.
[b]t-ratios (in absolute value) are beneath marginal effect estimates.
^ Significant at the .05 level.
* Significant at the .10 level.

Table 4: Marginal Effects Estimates for the Sellers' Quality Split by Dealer Type[a,b,c]

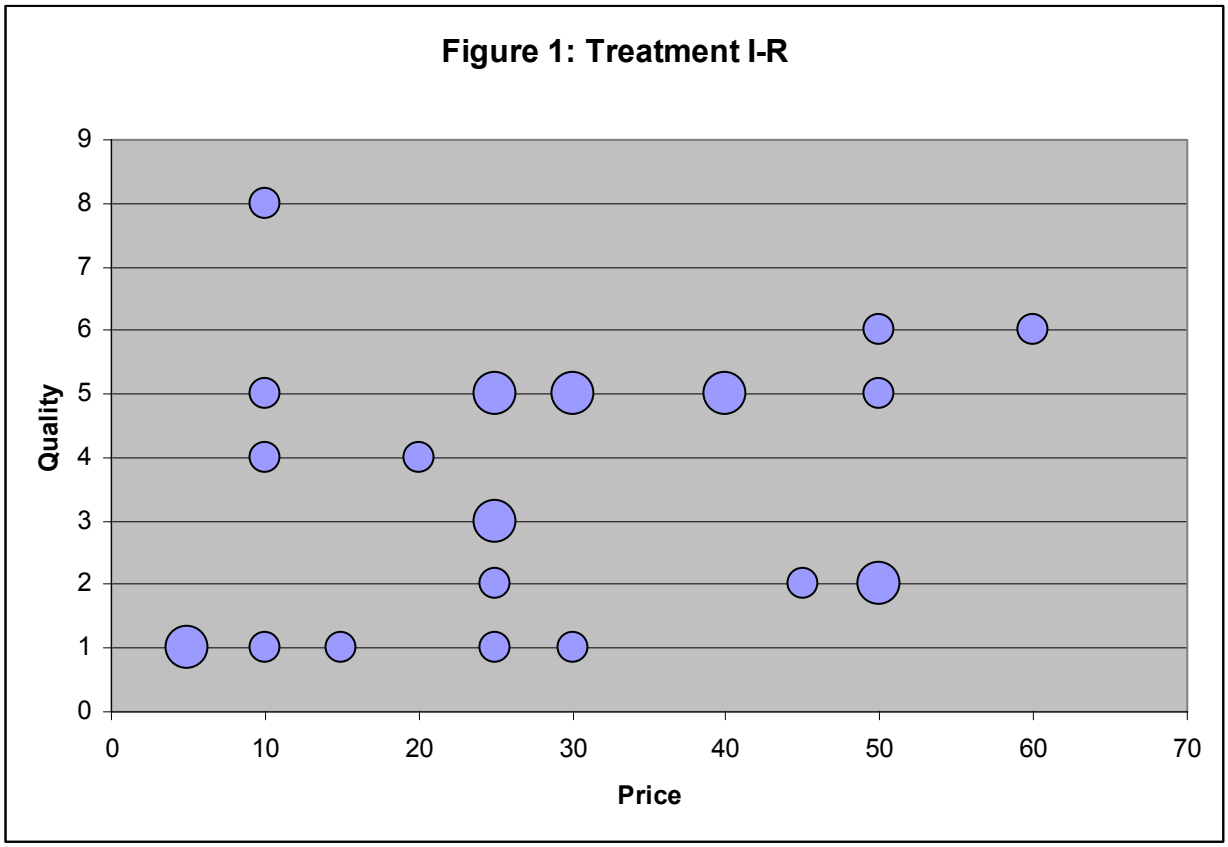
	Treatment Type								
Variable	**III_L**	**III_N**	**IV-NG_L**	**IV-NG_N**	**IV-AG_L**	**IV-AG_N**	**IV-G_L**	**IVG_N**	**IV-P_L**
Price	0.03^	0.004	0.002	-0.005	0.04^	0.003	0.04^	0.003	0.04^
	(8.6)	(0.7)	(0.2)	(0.5)	(2.1)	(0.3)	(2.7)	(0.1)	(4.8)
Constant	0.6^	0.6^	1.6^	1.8^	1.7^	1.5^	1.8^	1.8*	1.8^
	(4.1)	(4.6)	(5.0)	(5.2)	(5.2)	(4.6)	(5.0)	(1.7)	(10.0)
θ	$0.31^	$0.01	$0.02	-$0.006	$0.32	$0.02	$0.42	$0.03	$0.35^
	(5.2)	(0.5)	(0.4)	(0.5)	(1.4)	(0.6)	(1.5)	(0.1)	(2.1)
Person Random Effects	YES	YES	YES	YES	YES	YES	YES	YES	YES
N	70	30	36	24	30	24	20	16	50

[a]Dependent variable is the sellers' product quality given to the buyer. IV-P_L pools IV-AG_L and IV-G_L data. θ is the monetary gift exchange estimate, computed as $\partial v(q)/\partial P$.
[b]t-ratios (in absolute value) are beneath marginal effect estimates.
[c]"L" ("N") after treatment type denotes regression with "local" ("non-local") dealer data only.
^ Significant at the .05 level.
* Significant at the .10 level.

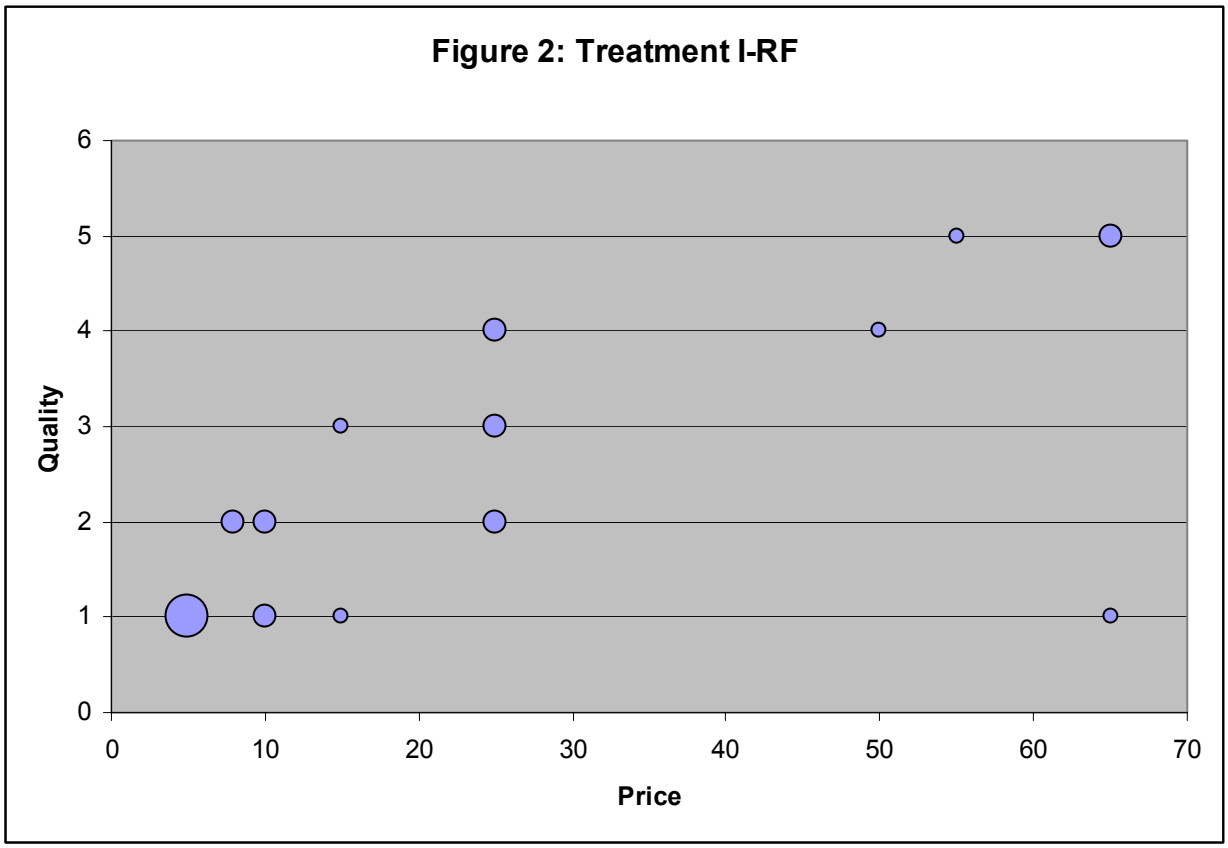
Table 5. Results Summary—Product Quality Claims

	Treatment III	Treatment IV-NG	Treatment IV-AG	Treatment IV-G
Overall	Claims: 53/100 Quality claim: 3.9(0.7) Delivered quality: 2.7(1.1) Delivered promised quality or above: 15/53	Claims: 36/60 Quality claim: 3.8(0.6) Delivered quality: 2.8(0.6) Delivered promised quality or above: 8/36	Claims: 24/54 Quality claim: 4.2(0.5) Delivered quality: 2.9(0.9) Delivered promised quality or above: 4/25	Claims: 17/36 Quality claim: 4.2(0.6) Delivered quality: 3.1(1.1) Delivered promised quality or above: 4/17
Local dealers	Claims: 27/70 Quality claim: 3.9(0.7) Delivered quality: 3.4(1.1) Delivered promised quality or above: 12/27	Claims: 22/36 Quality claim: 3.9(0.5) Delivered quality: 2.8(0.6) Delivered promised quality or above: 4/22	Claims: 7/30 Quality claim: 4.1(0.3) Delivered quality: 3.9(0.4) Delivered promised quality or above: 4/7	Claims: 4/20 Quality claim: 4.3(1.0) Delivered quality: 3.8(0.5) Delivered promised quality or above: 2/4
Non-local dealers	Claims: 26/30 Quality claim: 4.0(0.7) Delivered quality: 2.0(0.6) Delivered promised quality or above: 3/26	Claims: 14/24 Quality claim: 3.7(0.6) Delivered quality: 2.8(0.6) Delivered promised quality or above: 4/14	Claims: 17/24 Quality claim: 4.3(0.6) Delivered quality: 2.5(0.6) Delivered promised quality or above: 0/18	Claims: 13/16 Quality claim: 4.2(0.4) Delivered quality: 2.9(1.2) Delivered promised quality or above: 2/13

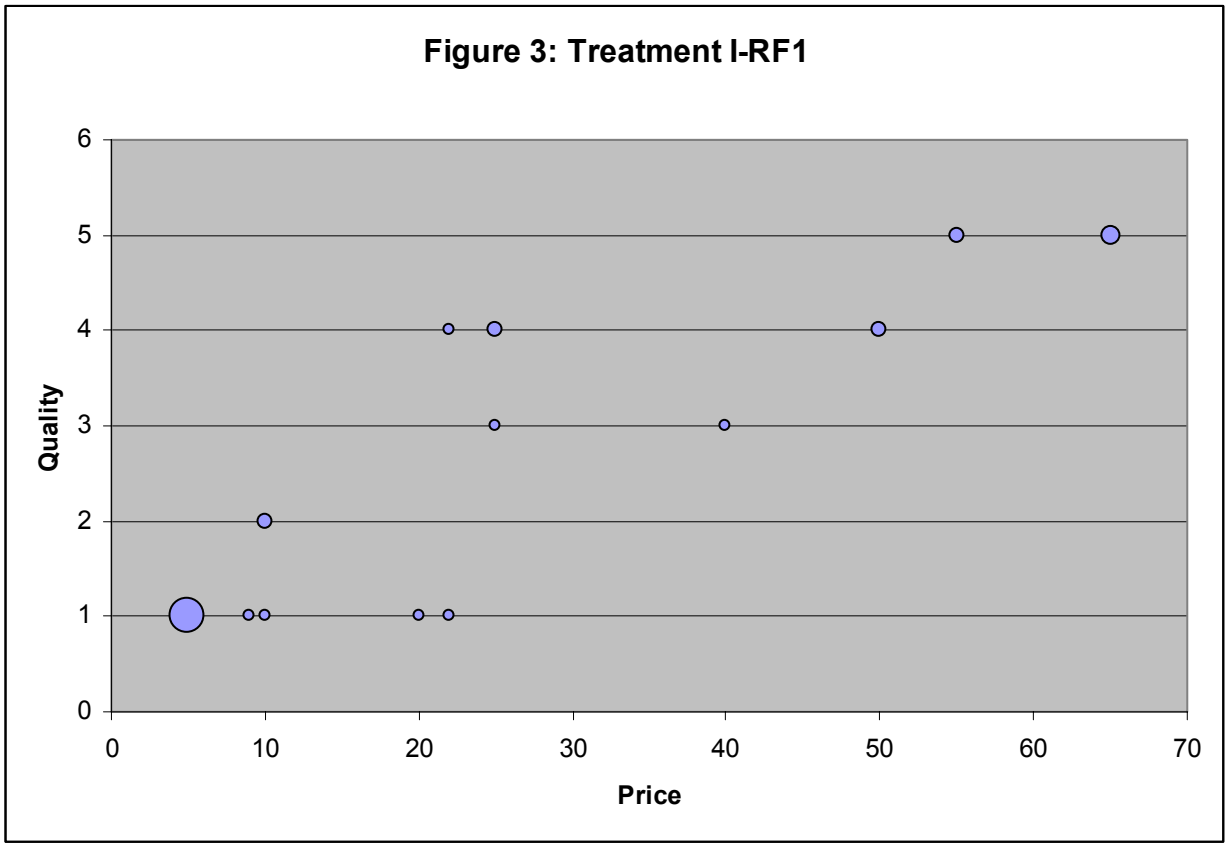
Notes: Each cell represents summary statistics from one unique treatment. For example, row 1, column 1, denotes that in the pooled Treatment III data, 53 of 100 dealer observations involved a claim of product quality. The average claim was 3.9 (PSA 8.9) and the average delivered product quality was 2.7 (PSA 7.7). Standard deviations are in parentheses beside means.



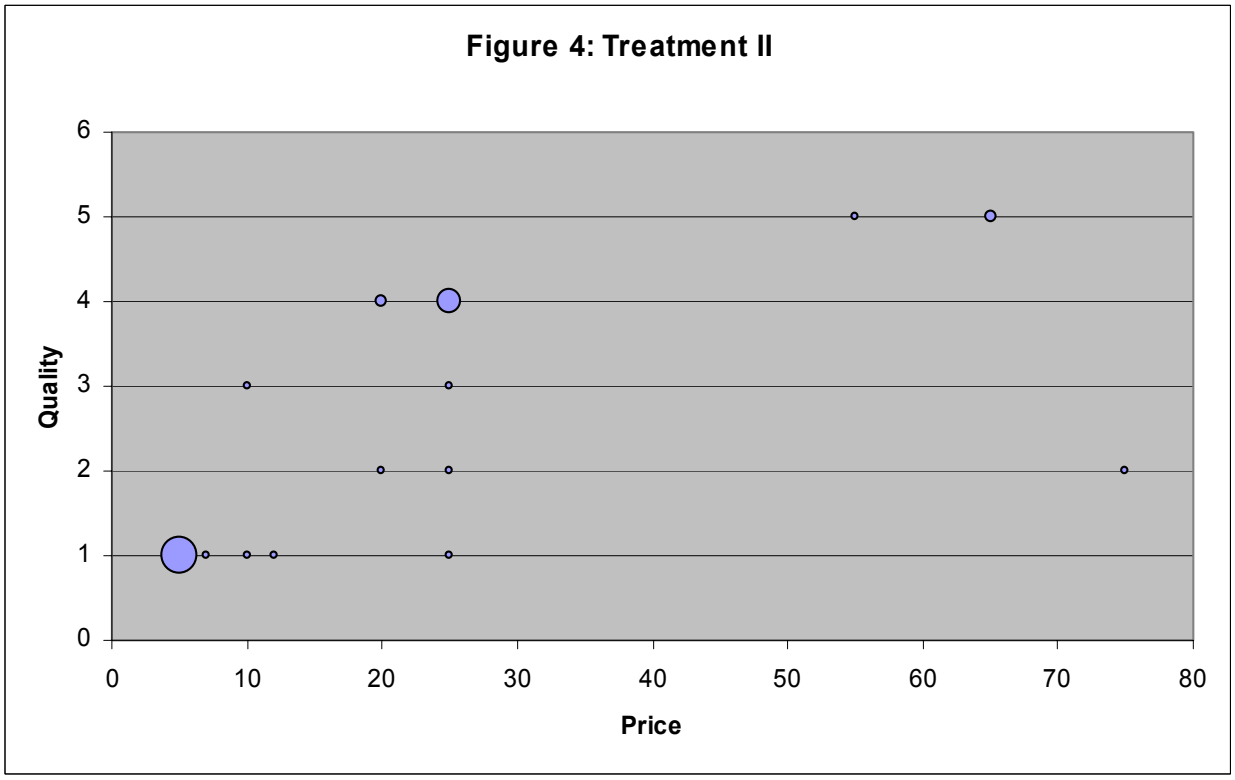
Note: Larger-sized circles indicate a greater number of observations occur at that point.



Note: Larger-sized circles indicate a greater number of observations occur at that point.



Note: Larger-sized circles indicate a greater number of observations occur at that point.



Note: Larger-sized circles indicate a greater number of observations occur at that point.

Figure 5: Price/Quality Relationship for Local Dealers

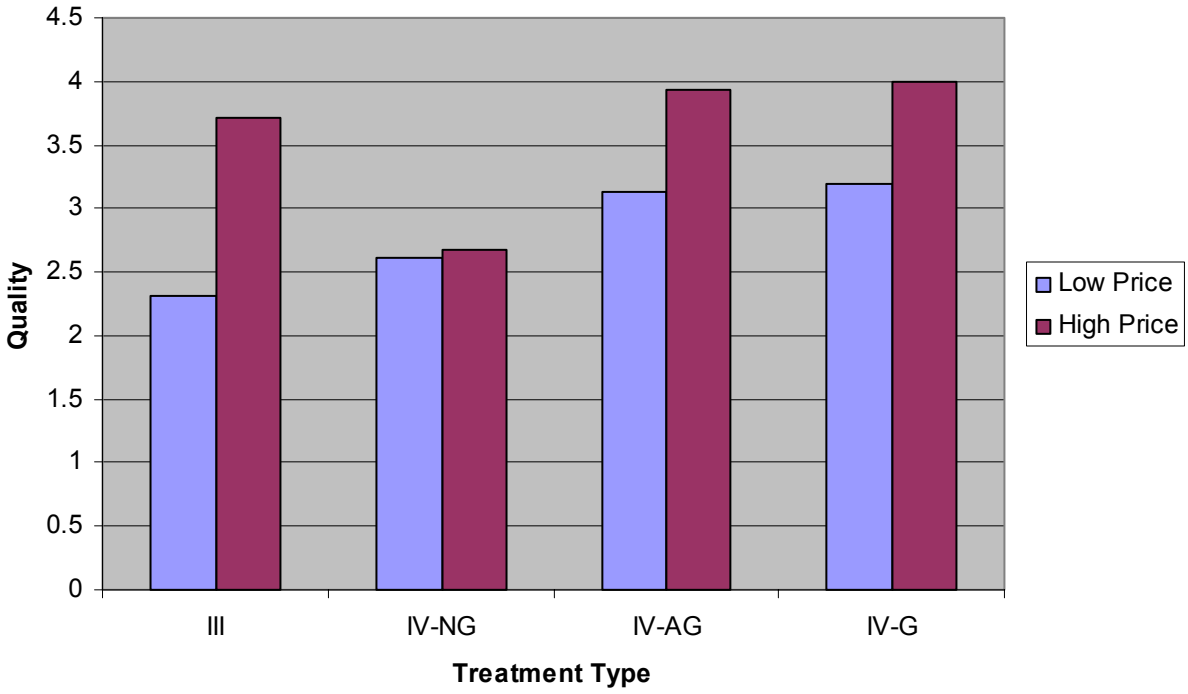


Figure 6: Price/Quality Relationship for Non-Local Dealers