Chapter 4

Group Reputations and Stereotypes as Contract Enforcement Devices

Incomplete contracts are frequently observed despite the well-known incentive distortions they create. For example, if a product's quality is not verifiable, sellers have a clear incentive to deliver lower-quality goods. Rational buyers recognize this incentive and adjust their demand accordingly. The resulting transaction is often Pareto dominated by the 'cooperative' outcome of high quality goods sold at higher prices. Despite these difficulties, cooperative market interactions continue to take place in the absence of complete, verifiable contracts; buyers often trust sellers to deliver a high quality product and sellers often respond in kind.

Economic theory has struggled to explain the success of the marketplace in the face of enforcement difficulties. An appealing argument is that repeated interactions act as an enforcement device when the cost of damaging a valuable long-term relationship outweighs the immediate benefit of poor performance, as in the models of Klein & Leffler [57] or MacLeod & Malcomson [65]. By adding a small probability of agents being unconditionally cooperative, Kreps *et al.* [59] show how false reputation-building by selfish agents can lead to full cooperation in early periods of the finitely repeated prisoners' dilemma. These models implicitly require that trading partners' identities be known in order for reputation-building to occur.

Experimental studies show, however, that cooperation can emerge even when interactions are anonymous. In tests of moral hazard in the labor market by Ernst Fehr and many others (see Fehr *et al.* [34] and [35], Charness [16], Fehr & Falk [32], Charness *et al.* [17], Gächter & Falk [39], and Hannan *et al.* [44], among others), wages and effort levels are observed to be substantially higher than the stage game equilibrium prediction even though transactions are anonymous. Furthermore, these studies show little evidence of reversion to the equilibrium in the final periods. Therefore, the authors conclude that fairness norms (such as a natural preference for 'gift exchange') solve the moral hazard problem, not reputation-building.

Not all experimental studies confirm these results. Lynch et al. [63], Engelmann & Ortmann [29], and Rigdon [84] find behavior consistent with, or converging toward, the stage game equilibrium. Even some studies purporting the existence of fairness preferences include some sessions with strong end-game effects, as in Fehr *et al.* [35] and Riedl & Tyran [83]. These apparently contradictory results leave open the question of what forces are at work to offset the shirking incentive.

The current chapter provides three novel results. First, in a direct replication of Fehr *et al.* [34], high wages and efforts observed in early periods collapse dramatically to the stage game equilibrium in the last period. Second, a group reputation (or 'stereotyping') model is developed that explains this behavior. Third, a new experimental environment is tested in which individuals have insufficient incentives to maintain the group's reputation. As predicted by the model, actual subjects play the stage game equilibrium in every period. Furthermore, the results of many previous experimental papers are consistent with the model's predictions.

The stereotyping model works as follows:

Assume, *à la* Kreps *et al.* [59], that some percentage of workers are unconditional cooperators whose efforts are always positively correlated with their wage. If firms believe that worker types are perfectly correlated – even if that belief is empirically unsupported – then a single defection by one worker leads to the belief that all workers are 'selfish,' destroying the reputation of the entire group and causing low wages in all subsequent periods.[1] Under the payoff structure used by Fehr *et al.* [34], selfish workers have an incentive to maintain a group reputation until the very last period (unless firms are very certain *a priori* that the workers are selfish.) By changing the payoff structure, one can eliminate the existence of such group reputation equilibria. Indeed, this is the phenomenon observed in a new set of experiments.

The assumption of stereotyping behavior, though irrational, is a well documented phenomenon in the social psychology literature. People use stereotypes to economize on cognitive resources in the processes of evaluating and recalling information about other individuals (McGarty et al. [70, p. 3–5]). Effectively, stereotyping is an inexpensive internal reputation management system. By assigning attributes to groups rather than individuals, decision makers can easily estimate the attributes of individ-

[1] This is similar in spirit to the contagion mechanism of Kandori [54].

ual group members. The danger of this cognitive shortcut is, of course, potentially harmful inaccurate beliefs about individuals.

The following section formally introduces the gift exchange market under various payoff structures and explores the equilibrium predictions of the stage game. Sections 4.2 and 4.3 describe experimental sessions in which high wages and effort are observed only under certain conditions, though all sessions show strong signs of repeated game considerations. Section 4.4 simplifies the structure of the gift exchange market for game theoretic analysis and Subsection 4.4.2 formally introduces the stereotype-reputation model, along with evidence that such a model has solid foundations in the field of social psychology. Subsection 4.4.3 compares the predictions of the stereotype-reputation model to past gift exchange market experiments, and Section 4.5 concludes the chapter.

4.1 The Gift-Exchange Market

A single play of a gift exchange market (GEM) has the following structure:

A finite set $\mathcal{J}$ of firms and a finite set $\mathcal{I}$ of workers participate in a two-stage posted-price labor market. Let $J = |\mathcal{J}|$ and $I = |\mathcal{I}|$, with $I > J$. In the first stage, each firm $j \in \mathcal{J}$ posts at most one wage offer w_j in the market. The set of allowable offers is given by $\mathcal{W} \cup \{\phi\}$, where $\mathcal{W} \subseteq \mathbb{R}$ is a compact set of non-negative wage offers and ϕ represents no wage offer. Workers $i \in \mathcal{I}$ may accept any outstanding offers w_j in the market at any time. After w_j has been accepted by some worker, firm j may not post any further wage offers, so each firm may hire at most one worker. After a fixed amount of time (or after all firms have had a wage offer accepted,) the first stage ends and unmatched agents earn zero profit.

In the second stage, each worker that accepted a wage offer selects an effort level e_i from a linearly ordered set $\mathcal{E}$ such that $\inf \mathcal{E} \in \mathcal{E}$. Let $\chi_i(w) \in \{0,1\}$ denote whether or not worker i accepts the given wage offer w and let $e_i(w) \in \mathcal{E}$ denote the effort level chosen. Note that by making workers the second mover, effort choices are quite naturally dependent on wages.

The monetary payoffs realized by each firm j and worker i are given by π and u, respectively, each mapping strategy pairs from $(\mathcal{W} \cup \{\phi\}) \times (\{0,1\} \times \mathcal{E})$ into $\mathbb{R}$ such that unmatched agents receive zero profit. The functions π and u are identical across agents and are common knowledge. Assume u is monotone decreasing in e and increasing in w, and π is monotone increasing in e and decreasing in w.

4.1.1 Stage Game Equilibrium

Define $e_{\min} = \inf \mathcal{E}$ and $w^* = \inf \{w \in \mathcal{W} : u(w,(1,e_{\min})) \geq 0\}$. Here, $e_{\min}$ is the minimal effort choice and w^* is the reservation wage at $e_{\min}$. Without loss of generality, assume that $e_i(w) = e_{\min}$ whenever $\chi_i(w) = 0$ since effort choices are irrelevant when no wage is accepted. In the second stage of the one-shot game, consider the strategy for each worker i given by

$$(\chi_i^*(w), e_i^*(w)) = \begin{cases} (1, e_{\min}) & \text{if} \quad w \geq w^* \\ (0, e_{\min}) & \text{if} \quad w < w^* \end{cases},$$

where $e_{\min}$ is chosen in response to any acceptable wage offer. This is the workers' unique dominant strategy since u is decreasing in e and $u(w,(1,e)) < u(w,(0,e_{\min})) = 0$ for each $e \in \mathcal{E}$ when $w < w^*$. Note that while the choice of whether to accept a given wage offer is dependent upon w, the choice of $e_{\min}$ is not.

In the current set of environments, $\pi(w^*,(1,e_{\min})) > 0$, so the firms have an incentive to participate. The unique subgame perfect equilibrium outcome is for each firm j to offer $w_j = w^*$ and for every worker i to accept w^* with an effort level of $e_{\min}$.[2] Since $I > J$, involuntary unemployment will still result.[3]

4.1.2 Three Specifications

Three specifications of the GEM are tested experimentally. Each varies in the functional form of agent payoffs and in the information feedback conditions, although the game forms are identical across specifications. In all three treatments, $I = 9$, $J = 6$, $\mathcal{E} = \{1, 2, \ldots, 10\}$ and $\mathcal{W} = \{5, 10, 15, \ldots\}$. Utilities are chosen such that $w^* = 30$, so the stage game equilibrium always predicts a wage of 30 and an effort choice of $e_{\min} = 1$.

[2] There exists a Pareto-dominated 'no-trade' Nash equilibrium to this game in which all J firms choose to make no wage offers ($w_j = \phi$ for all $j \in \mathcal{J}$) and all I workers choose to reject all wage offers ($\chi_i(w) \equiv 0$ for all $i \in \mathcal{I}$.) Unless otherwise indicated, further discussion of the stage game equilibrium will refer only to the subgame perfect equilibrium with full employment.

[3] In equilibrium, workers know that w^* will be the only wage offer in the market and will therefore accept w^* immediately. Allocation of firms to workers is assumed to be random in the situation of multiple simultaneous acceptances, so that the set of unemployed workers will be randomly selected in equilibrium.

4.1.3 Treatment 1: High MRS Ratio, Anonymous IDs (HRA)

The first variant of this game – denoted HRA – is the original gift exchange market studied by FKR. Here, the payoffs for matched firms and workers are given by

$$\pi_1(w,(1,e)) = (126-w)\left(\frac{1}{10}e\right) \tag{4.1}$$

and

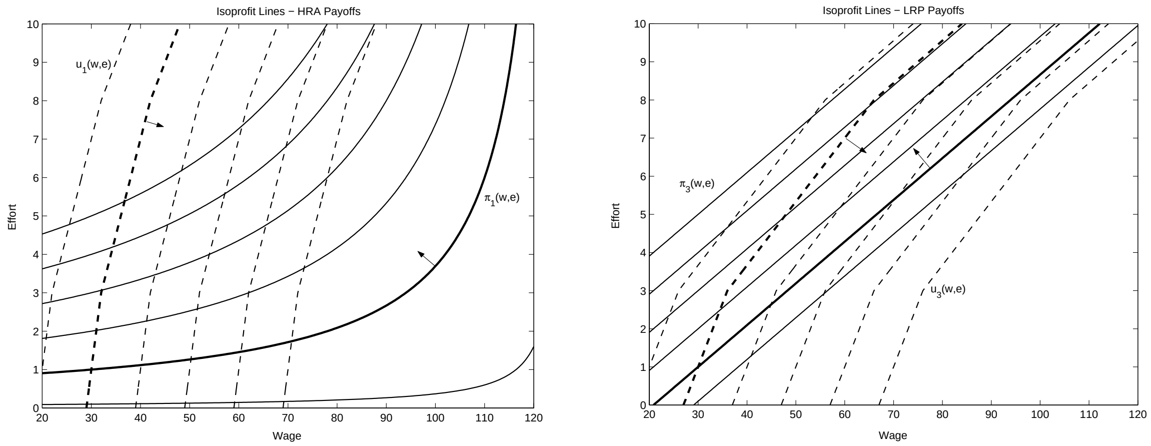
$$u_1(w,(1,e)) = w - 26 - c(e), \tag{4.2}$$

where the cost of effort is given by

$$c(e) = \begin{cases} -1+e & \text{if} \quad e \in \{1,2,3\} \\ -4+2e & \text{if} \quad e \in \{4,5,6,7\} \\ -12+3e & \text{if} \quad e \in \{8,9,10\} \end{cases} \quad .$$

In the experiment, π_1 and u_1 are denoted in francs which are then converted to dollars at a rate of 12 francs per dollar.

At the equilibrium strategy profile $(w^*, e_{\min})$, each firm's marginal rate of substitution is $(\partial\pi_1/\partial e)/(\partial\pi_1/\partial w) = -1$ and each worker's MRS is $-1/96$. The ratio of these values is quite high, indicating that each party in a transaction can have a dramatic effect on the payoffs of the counterparty at little cost to themselves. As strategies move farther from equilibrium, the cost of affecting the other agent's profit increases. This is clear from the graph of the level curves of π_1 and u_1 in the space $\mathcal{E} \times \mathcal{W}$, given in Panel (a) of Figure 4.1.



(a) **(b)**

Figure 4.1: Isoprofit lines for workers and firms in (a) HRA – the gift exchange market used by Fehr, Kirchsteiger & Riedl, and (b) LRP, the new design with quasilinear payoffs.

In HRA, wage offers are displayed for all agents to see, but the identity of the firm offering each wage is known only to the firms. Similarly, the acceptance of wage offers is public information, but the identity of the accepting worker is known only among the workers. Finally, the effort level decision of each matched worker is made after the market is closed and revealed only to the hiring firm. No other firms or workers observe this decision. Thus, it is impossible for firms to develop informative reputations about individual workers in this environment. In sessions using HRA, subjects are paid one U.S. Dollar for every 12 'francs' earned in the experiment.

4.1.4 Treatment 2: High MRS Ratio, Public IDs (HRP)

The second variant – HRP – alters the information structure and payoff conversion rates of HRA. First, all agents observe the player ID number associated with each wage offer and with each worker accepting any given wage offer. Second, effort level decisions are made immediately after a worker accepts a wage offer and this decision is posted (along with the worker's ID number) for all agents to observe. This not only provides information for the formation of individual reputations across periods, but also allows all agents to observe the realization of strategies chosen by each worker, given the accepted wage offer before the market closes. Finally, the conversion rate between experimental currency and actual payoffs is increased to four francs per dollar for the workers and nine francs per dollar for the firms so that consequences of strategy choices have increased saliency.

The payoff functions of HRP are identical to HRA, so that $\pi_2 \equiv \pi_1$ and $u_2 \equiv u_1$. Thus, the high ratio of the marginal rates of substitution is maintained.

4.1.5 Treatment 3: Low MRS Ratio, Public IDs (LRP)

The third variant, LRP, alters HRP to make the payoffs of both agents quasilinear in wages. The cost of effort function is tripled to reduce the disparity between the effect of a change in effort on workers and firms. Finally, a linear rescaling of the 'value of effort' to the firms is used to adjust payoffs for the experimental environment. These changes serve to reduce the equilibrium MRS ratio so that workers cannot dramatically affect the payoff of the firms without significant penalty to themselves, and vice-versa. The payoff functions for LRP are given by

$$\pi_3\left(w,(1,e)\right) = 126\left(\frac{11}{40}+\frac{2.9}{40}e\right) - w \tag{4.3}$$

and

$$u_3\left(w,(1,e)\right) = w - 26 - 3c\left(e\right). \tag{4.4}$$

The conversion rate of 12 francs per dollar is used for all subjects.

The ratio of marginal rates of substitution at the equilibrium is three in this design, as opposed to 96 in the previous treatments. Keeping the ratio greater than one guarantees that there still exists wage-effort pairs that Pareto dominate the equilibrium so that players still have an incentive to attempt cooperation. The level curves of the payoff functions are given in Panel (b) of Figure 4.1.

4.2 Experimental Design

The three designs were tested experimentally at the California Institute of Technology Laboratory for Experimental Economics and Political Science (EEPS) using Caltech undergraduate students recruited via e-mail. Subjects were randomly divided into two groups of six firms and nine workers, with each group separated into a different room. The instructions did not make reference to firms, workers, wages, or effort levels.[4] To avoid a labor market framing, subjects were labelled as buyers and sellers,

[4] For the HRA session, the instructions provided by Fehr et al. [34] were used. For HRP and LRP, the instructions were appropriately modified. Copies of the instructions are available upon request.

and their task was to post prices for a good in a market and choose a 'conversion rate' (rather than an effort level) that affected payoffs. The term 'conversion rate' is used in HRA and HRP to emphasize that sellers, by their choice of e, are choosing the percentage of $(126 - w)$ that their buyer will be paid. In LRP, the effort level choice can no longer be thought of as a conversion rate on firms' profits, so the generic name, 'X' was instead used in the instructions to identify this choice variable.

Telephone conversations between two experimenters was used as the means of transmitting information between rooms during the market stage of each period. All decisions were posted on the blackboard in both rooms.[5] When effort levels were not publicly viewable, the worker wrote his effort decision on an index card that was delivered to the appropriate subject in the other room. The first session (S1) consisted of HRA repeated over twelve periods. The second (S2) ran HRP for twelve periods, while the third and fourth sessions (S3 and S4) ran LRP for twelve periods. The fifth session (S5) was divided into two parts.[6] First, LRP was played for six periods. Immediately following, the same subjects read instructions and participated in HRA for six periods.[7] The treatment-switching design in S5 tests whether or not social norms or reputations developed in LRP affect behavior in HRA which can then be compared to behavior in S1. If behavior is substantially different between HRA and LRP within S5, then differences in the structure of the two markets apparently

[5] In later sessions, the market information was projected on a screen and transmitted by computer. This had no effect on the actual procedures of the experiment.

[6] Sessions were run on the following dates: S1 on 12/03/2002, S2 on 12/05/2002, S3 on 08/05/2004, S4 on 08/04/2004, and S5 on 01/14/2003.

[7] Although subjects were informed that they would participate in two different experiments, they were not given specific information or instructions about the second treatment until the conclusion of the first.

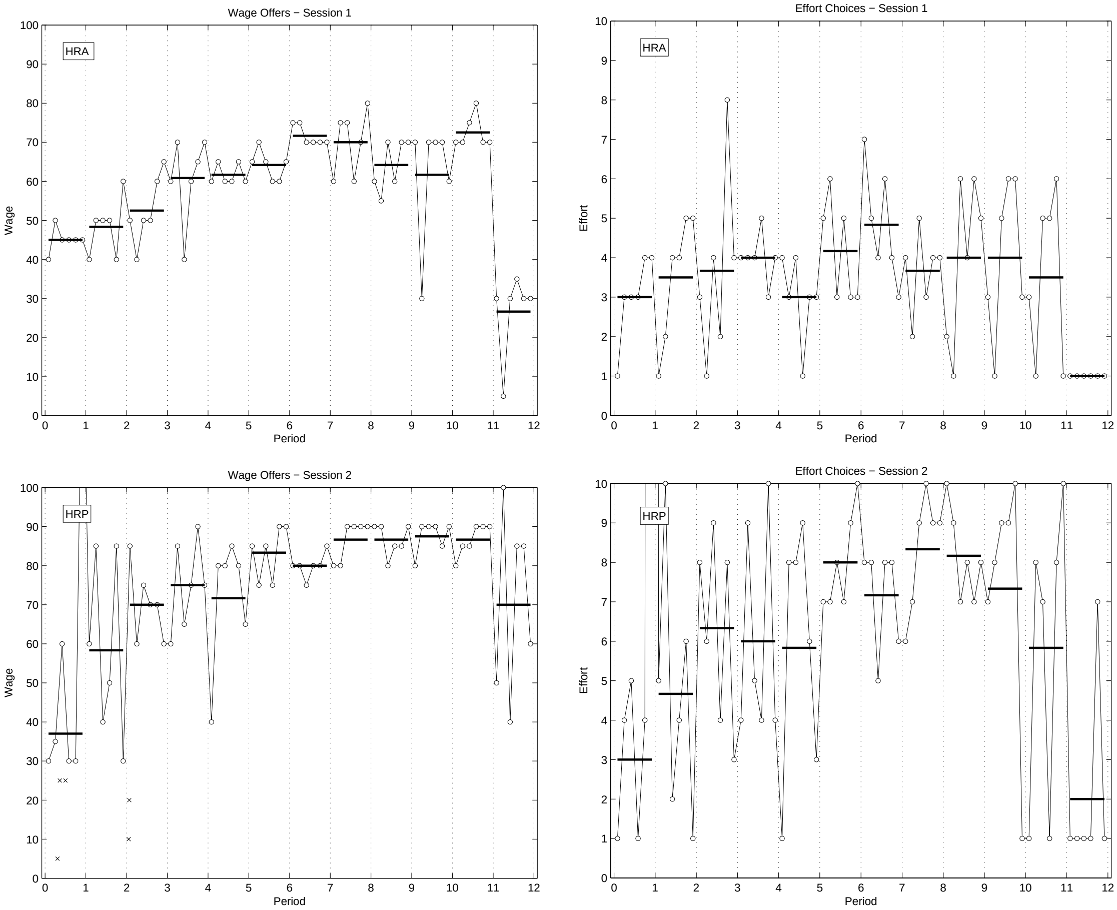
cause differences in behavior. Each session lasted between 90 minutes and two hours. In sessions S1 and S5, subjects earned an average of $35, while earnings in S2 were as high as $130 due to the reduced exchange rate. In S3 and S4, average earnings were around $25 because of the reduced cooperation in that treatment.

4.3 Experimental Results

See Figures 4.2, 4.3 and 4.4 for a complete representation of the data from the five experimental sessions.[8] These results show that effort does appear to be an increasing function of wages. In HRA and HRP, wage-effort pairs are well above equilibrium, but play converges to the stage game equilibrium in the final period. In LRP, players are unable to coordinate on high wage-effort pairs and the stage game equilibrium is observed across all periods. In HRA played after LRP, subjects are able to coordinate on high wage-effort pairs, indicating coordination is easier in the latter environment. These results are inconsistent with a pure fairness hypothesis, but due to the information structure of HRA, must also be inconsistent with individual reputation building. Thus, the results are indicative of a group reputation effect.

In previous experiments, a strong positive correlation between wages and efforts has been observed. This key wage-effort relationship is traditionally taken as evidence

[8] In session S4, two subjects acting as workers had not been matched with many wage offers in the first several periods and consequently, had accumulated very little earnings by the 7^{th} and 8^{th} periods. These subjects, informed that they would not have to pay their losses to the experimenter, began to accept the smallest possible wages and offer the highest possible effort in an attempt to create maximal wealth for the (anonymous) firms. After 4 such actions, one worker was removed from the experiment and the other immediately (and voluntarily) stopped participating. These four data points are removed from analysis, but likely affected beliefs in the market for the remainder of the session.



Wage Offers − Session 1

Effort Choices − Session 1

Wage Offers − Session 2

Effort Choices − Session 2

Figure 4.2: Wage and effort levels across time in sessions S1 (a replication of the Fehr, Kirchsteiger and Riedl (1993) experiment,) and S2 (the same design with individual reputations added.) Solid lines represent period averages and x's represent unaccepted bids.

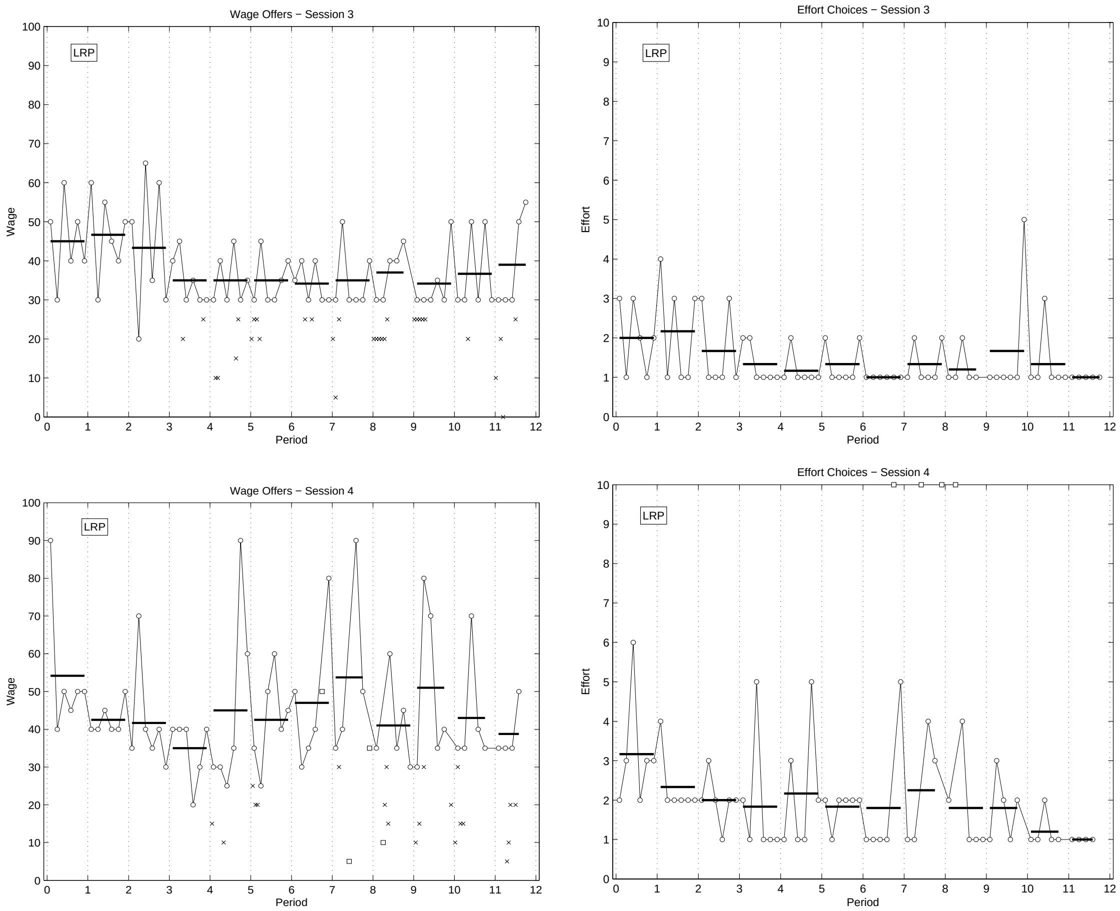


Figure 4.3: Wage and effort levels across time in sessions S3 and S4 (with quasilinear payoffs.) Solid lines represent period averages and x's represent unaccepted bids. Four data points in S4 (represented by squares) are removed from analysis. See footnote 8.

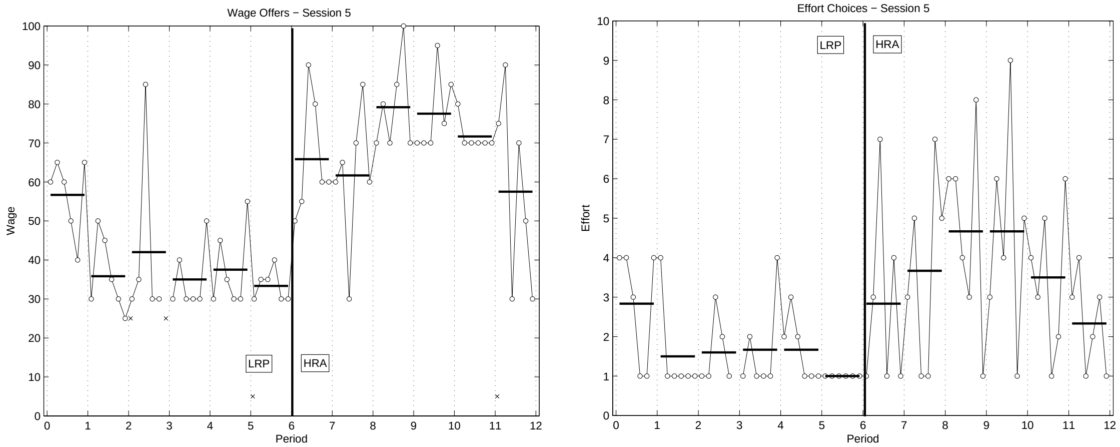


Figure 4.4: Wage and effort levels across time in session S5, switching from LRP to HRA after period 6. Solid lines represent period averages and x's represent unaccepted bids.

of reciprocal-minded subjects, but is also consistent with repeated game explanations in which selfish subjects either mimic reciprocal agents in early periods (a reputation effect) or use punishment strategies to enforce cooperation (a folk theorem effect.) Regardless of the cause, this correlation is also observed in the current set of experiments.

Session (Treatment)	RankCorr(w, e)	p-Value
S1 (HRA)	0.546	7.07×10^{-7}
S2 (HRP)	0.641	1.64×10^{-9}
S3 (LRP)	0.604	2.64×10^{-7}
S4 (LRP)	0.446	0.0001
S3 (LRP)	0.499	0.0023
S3 (HRA)	0.511	0.0015

Table 4.1: Spearman rank correlation coefficients between effort and wages for each treatment

Result 4.1 *In all treatments in all sessions, wages and efforts of matched firms and workers are positively correlated as predicted by both fairness and repeated game explanations.*

Support. Spearman rank correlation coefficients between wages and effort are calculated for each treatment. For each, the coefficient is estimated to be at or greater than 0.446 and significantly positive at the 0.5 percent level. The estimates of the individual coefficients are given in Table 51. ■

Although the wage-effort correlation is inconclusive about the strategies employed, the following set of results provide evidence in support of a group-reputation repeated game explanation.

Result 4.2 *In the replications of the FKR gift exchange market (HRA), strategies converge toward the stage-game equilibrium in the final period, providing evidence of repeated game strategies.*

Support. See Figures 4.2 and 4.4. In session S1, the average wage for periods 4 through 11 is 65.83. In period 11, the final wage offer of 70 was given a minimal effort level. In the final period (period 12), all wages except one and all efforts are at or below the equilibrium values, and the average wage was less than the equilibrium prediction.[9] In the latter half of session S5, average wages decline in the final periods and two wage offers appear at equilibrium levels in the last period, although the average wage offer only declines to 57.5. Similarly, average effort choices decline in

[9] One worker accidentally accepted a sub-equilibrium wage offer in his haste to participate in the market.

the final periods and two last-period effort choices are at the equilibrium prediction, although the average declines to only 2.33. ■

For any model of fairness to accommodate these results, agents' relative preference for fairness must be time-dependent. Furthermore, subject pool effects are significant in the last period since the HRA represents an exact replication of the Fehr *et al.* [34] experiment using different subjects. The percentage of purely fair-minded agents is apparently lower in the population used for the current study, although high wage-effort choices are observed in early periods across subject pools.

Result 4.3 *In the first 11 periods of the 12-period replication of the FKR experiment (HRA), the average wage increases in time, while average effort does not significantly change.*

Support. Each transaction is numbered chronologically from 1 to 72, with six transactions per period. The rank correlations between transaction numbers and wages and between transaction numbers and effort are estimated for the first 66 transactions (11 periods) using Spearman rank correlation coefficients. Rank correlation between wages and transaction number is estimated at 0.6905 with a 2-sided p-value of 1.4×10^{-10}, while correlation between effort and transaction number is estimated at 0.1711 with a p-value of 0.1696. Rank correlations are also estimated using period number as a proxy for time instead of transaction number, yielding very similar results. ■

That wages increase in time is not predicted by either theory under consideration, but it is not necessarily inconsistent with a reputation-building equilibrium. On the other hand, fairness explanations are time-independent, suggesting that pure fairness concerns are inadequate to explain behavior.

The following results indicate that the ability of players to achieve outcomes that Pareto dominate the stage game equilibrium is not robust to the payoff specifications.

Result 4.4 *In LRP, the minimum effort level is played more often than all other strategies combined, and the effort level regresses to the stage game equilibrium strategy in the final two periods, suggesting that behavior is highly sensitive to the market's payoff structure.*

Support. Of the 169 effort decisions in LRP treatments, 60.4 percent are at $e = 1$. Over 91 percent of all observations are at effort levels 1, 2, or 3. Only 5 effort choices are observed at $e \geq 5$. In the penultimate period, the minimal effort is observed in 12 of the 17 transactions, with an average effort of 1.411. In the final period, *every one* of the 15 sellers selects $e = 1$. ■

Result 4.5 *In LRP, the firm's stage game equilibrium strategy ($w = 30$) is the modal observation and the frequency of this strategy increases with time.*

Support. The subgame perfect strategy of $w = 30$ occurs in 32 percent of the 169 accepted wage offers in treatment LRP – more often than any other strategy is used. Wage offers of $w \leq 40$ constitute 68 percent of all observations and 86.4 percent of accepted wage offers are no greater than 50. ■

If fairness considerations are independent of payoff specifications, then fair-minded workers will always be willing to reciprocate in all gift exchange environments. This is not observed in the LRP experiment. The following result indicates that the difference in environment is responsible for the differences in behavior.

Result 4.6 *In the treatment-switching session (S5), average wages and effort levels increase after switching from LRP to HRA, indicating that changing the payoff structure back to that of FKR induces the same group of subjects to choose high wages and effort.*

Support. To avoid problems with non-stationarities in the time series, each wage and effort from LRP is compared to the wage and effort from HRA with the same time identifier (or "transaction number"). These differences are analyzed using a Wilcoxon signed rank sum test. For wages, the HRA values are significantly greater than those from LRP, with an estimated z-statistic of 4.693. Similarly, HRA effort choices are significantly greater with a z-statistic of 3.652. Thus, significance in both cases is better than 0.015 percent. ■

This result is obvious from Figure 4.4. It is clear that the change in behavior immediately follows the change of the market in the seventh period. The fact that the same group of subjects generates two very different sets of data in two similar markets played consecutively implies that the difference in behavior is due to differences in the two markets. Therefore, the structure of HRA appears much more conducive to high wage-effort pairs than that of LRP.

Result 4.7 *In a replication of FKR's experiment with full information about past strategy choices (HRP), wages and effort are significantly greater than in the original design (HRA).*

Support. Again, wages and effort were paired between sessions according to their time identifier and a Wilcoxon signed rank sum test was performed in the differences. Wages are significantly higher in HRP, with a z-statistic of 5.925. Effort is also significantly higher in HRP, with a z-statistic of 5.401.[10] ■

Recall that the only differences between HRP and HRA are that the worker ID numbers and effort levels are publicly displayed with each transaction and the payoffs are increased to 4 francs/dollar for the workers and 9 francs/dollar for the firms. Thus, individual reputations and/or increased payoffs lead to more cooperation. Comparing HRP to LRP, however, reveals that cooperative behavior is sensitive to the payoff specifications, and not to the information structure of the game. This is consistent with a reputation equilibrium hypothesis in which firms are more likely to engage in trusting behavior if the relative payoff is greater.

4.4 A Reputation Model With Stereotypes

The results of Section 4.3 indicate that the stage game equilibrium prediction obtains in the LRP environment, but not in HRA or HRP until the final period. However, the remarkable reversion to the stage game prediction in the final period of both HRA and HRP indicates that a simple model of fairness is inadequate to explain the high effort and wages observed in early periods. The final period crash is reminiscent of a repeated game sequential equilibrium with reputations, as in Kreps *et al.* [59] and Kreps & Wilson [60].

[10] That this difference is more significant than that documented in the previous result is also a consequence of the larger sample size.

As Proposition 4.1 will demonstrate, a standard sequential equilibrium cannot explain the observed behavior since worker decisions are not linked to ID numbers. However, the psychology literature on categorical thinking and stereotyping indicates that it may be appropriate to assume that firms erroneously believe worker types are correlated. Proposition 4.2 shows that under such an assumption, the observed behavior in *all three* sessions is consistent with a sequential equilibrium explanation. Furthermore, the results of many past studies are also explainable in this framework.

4.4.1 The Basic Framework

To apply more complex game theoretic arguments to the gift exchange market, a simplified version of the GEM is developed. This 'mini-GEM' consists of a stage game repeated over T periods. In each period t, let $\mathcal{W} = \{\underline{w}, \overline{w}\}$ with $\overline{w} > \underline{w}$, and let $\mathcal{E}(\overline{w}) = \{\underline{e}, \overline{e}\}$, and $\mathcal{E}(\underline{w}) = \{\underline{e}\}$ with $\overline{e} > \underline{e}$, so that allowable effort choices are a function of the wage.[11] A fraction J/I of the workers are chosen with equal probability each period and matched with a wage offer. Let $j(i,t)$ denote the firm matched to worker i (if any) in period t, and $i(j,t)$ denote the worker matched to firm j. Define $w_{j,t}$ and $e_{i,t}$ to be the wage and effort in period t of firm j and worker i, respectively. Unmatched workers have no strategy choice and receive zero payoff.

Workers' types ($\boldsymbol{\theta}_i$) are selected from the set $\Theta = \left\{\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}\right\}$, where $\underline{\boldsymbol{\theta}}$ represents a 'selfish' worker and $\overline{\boldsymbol{\theta}}$ represents a 'reciprocal' or 'fair' worker. Worker payoffs for

[11] The restriction of $\mathcal{E}(\underline{w}) = \{\underline{e}\}$ is for simplicity of exposition. As will be apparent, allowing $\mathcal{E}(\underline{w}) = \{\underline{e}, \overline{e}\}$ does not alter the equilibrium analysis since selecting a high effort in response to a low wage can be assumed to perfectly signal that the agent is *not* reciprocal, thus causing reduced current period payoffs *and* reduced continuation payoffs. Furthermore, Clark & Sefton [23] experimentally demonstrate that in a sequential prisoners' dilemma, $\overline{e}$ is played in response to $\underline{w}$ in less than 5% of decisions.

type $\underline{\boldsymbol{\theta}}$ in each stage t are given by the function $u\left(w_{j(i,t),t}, e_{i,t}|\underline{\boldsymbol{\theta}}\right)$ whose values exactly match the monetary payoff specifications of the game given $\chi = 1$. Instead, type $\overline{\boldsymbol{\theta}}$ workers receive payoffs given by

$$u\left(w_{j(i,t),t}, e_{i,t}|\overline{\boldsymbol{\theta}}\right) = \begin{cases} -1 & \text{if} \quad \left(w_{j(i,t),t}, e_{i,t}\right) = (\overline{w}, \underline{e}) \\ 0 & \text{otherwise} \end{cases}$$

Thus, selfish workers receive utility for only their monetary payoffs and reciprocal workers (type $\overline{\boldsymbol{\theta}}$) are penalized whenever they fail to reciprocate. Time discounting is ignored since experimental subjects are paid for all decisions at the end of the session.

Figure 4.5 shows the extensive form of one period of the mini-GEM for one matched pair of players with $\mathcal{W} = \{30, 100\}$ and $\mathcal{E} = \{1, 10\}$, using either the HRA or LRP payoffs. It is clear that $(\underline{w}, \underline{e})$ is the unique Nash equilibrium profile of either stage game when the worker is selfish ($\boldsymbol{\theta}_i = \underline{\boldsymbol{\theta}}$). However, if the firm believes that $\overline{e}$ will obtain with sufficiently high probability, then $\overline{w}$ is the best response. This depends on the firm's subjective probability that the worker is a reciprocating agent and the relative payoffs for cooperation ($\overline{w}$,$\overline{e}$), defection ($\overline{w}$,$\underline{e}$), and equilibrium ($\underline{w}$,$\underline{e}$).[12]

In each period of the repeated game, a subset of J workers is randomly selected to be paired with the J firms, while $I - J$ workers remain 'unemployed' for the period and receive zero payoff. A sequential equilibrium of this game is defined as a pairing of a strategy profile and a system of beliefs such that each agent is playing optimally at every information set, given the strategies of others and the system of

[12] The terms cooperation, defection, and equilibrium are taken from the prisoners' dilemma literature since this game is effectively a sequential prisoners' dilemma.

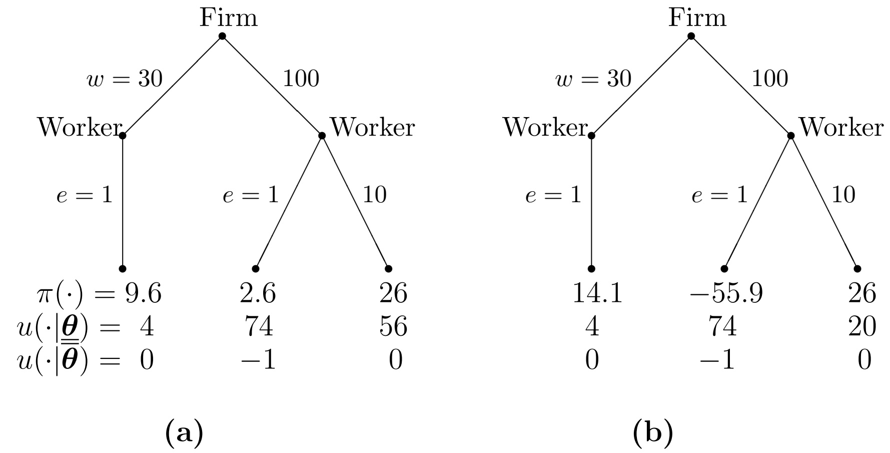


Figure 4.5: **(a)** The HRA mini-game, and **(b)** the LRP mini-game, assuming $\mathcal{W} = \{30, 100\}$ and $\mathcal{E} = \{1, 10\}$.

beliefs and the beliefs at each information set are derived from previous beliefs and action probabilities in accordance with Bayes' Law.

For notational simplicity, let $A = \pi(\underline{w}, \underline{e}) - \pi(\overline{w}, \underline{e})$, $B = \pi(\overline{w}, \overline{e}) - \pi(\overline{w}, \underline{e})$, $C = u(\overline{w}, \underline{e}) - u(\overline{w}, \overline{e})$, and $D = u(\overline{w}, \underline{e}) - u(\underline{w}, \underline{e})$. Here, A/B serves as a measure of how tempted the firms may be to gamble on the worker types by offering a high wage; if A/B is near zero, then offering a high wage is more appealing. Similarly, C/D is a measure of the workers' temptation to defect.

Under the assumption that worker types are independent and $e_{i,t}$ is only observed by $j(i,t)$, only the beliefs of firm $j(i,t)$ are changed by i's choice in period t. Let $p_{j,t}$ denote firm j's probability in period t that the randomly assigned worker will have type $\overline{\boldsymbol{\theta}}$. If $p_{j,t+1}$ is not substantially lower than $p_{j,t}$ when $e_{i(j,t),t} = \underline{e}$, then firm j's strategy in $t + 1$ will be the same, regardless of $e_{i(j,t),t}$. If this is the case and $i(j,t)$ is selfish, then $\underline{e}$ will be chosen. Furthermore, tight restrictions on C and D are needed to guarantee that selfish workers would indeed prefer to maintain a false reputation, giving the following result:.0

Proposition 4.1 *If worker decisions are anonymous and private, worker types are independent, and $C/D > 1/I$, then there does not exist a reputation equilibrium of the T-period repeated mini-GEM in which all workers choose $\overline{e}$ with probability 1 in all periods $t < T$ and all firms choose $\overline{w}$ in every period.*

The proof of this proposition follows from the arguments used in the proof of Proposition 4.2 which appears in the chapter appendix (Section 4.6.) Note that the equilibrium described in this proposition describes observed behavior in HRA. However, letting $\overline{w} = 100$, $\underline{w} = 30$, $\overline{e} = 10$, and $\underline{e} = 1$, the weak condition on C and D is satisfied for both HRA and LRP, so no pure strategy reputation equilibrium can exist in these two mini-games. Thus, for the experimental data to be explained by this model, either there exists a mixed strategy equilibrium that predicts frequent high effort levels, or firms believe that worker types are instead, highly correlated. Mixed strategy reputation equilibria are unlikely to exist, and if they do, are unlikely to generate high effort levels through many periods.[13] Therefore, a model of perceived

[13] To see this, let $q_{i',s}$ be the probability that agent i' chooses $\overline{e}$ in response to $\overline{w}$ in each period s, and let $q_{i',s} = 1$ if i' did not participate in period s. Note that if i' selects $\underline{e}$ in response to $\overline{w}$ in some period $s < T-1$, then the beliefs of the firms in period $T-1$ about agent i', (denoted $p_{i',T-1}$), will be zero. Given some worker $i \in \mathcal{I}$, define the average belief assigned to the other workers going into the final period to be

$$\bar{p}_{-i,T} = \frac{1}{I-1} \sum_{\substack{i' \in \mathcal{I} \setminus \{i\}: \\ p_{i',T-1} > 0}} \frac{p_1}{p_1 + (1-p_1)\prod_{s=1}^{T-1} q_{i',s}}.$$

In period T, worker i will choose $\overline{e}$ in response to $\overline{w}$ only if

$$\frac{A}{B} - \frac{1}{I} < \frac{I-1}{I}\bar{p}_{-i,T} < \frac{A}{B}.$$

type correlation is examined.

4.4.2 The Model With Stereotypes

Although there exists experimental support for reputation equilibria in repeated games (for example, see Camerer & Weigelt [12], Neral & Ochs [76], and Andreoni & Miller [3]), the added assumption that firms view workers' types as highly correlated is apparently irrational. For example, Fehr *et al.* [35] and Fehr & Falk [32] find support for heterogenous player types in gift exchange markets, so rational, well-informed subjects should not expect homogeneity or strong correlation. On the other hand, McEvily *et al.* [69] show that subjects make inferences about the trustworthiness of opponents based on whether or not *different* opponents were trustworthy in early periods. Furthermore, when the group of opponents is chosen according to some unrelated criterion, the effect becomes even more pronounced. Thus, decision makers use past behavior to make inferences about the future behavior of others, especially when there is any reason to think those individuals are part of a group; subjects perceive correlation between their anonymous competitors.

The existing social psychology literature also supports the claim that people in social situations infer more correlation than is warranted – a phenomenon known as 'illusory correlation.' By design, the gift exchange market separates firms and workers into groups before the experiment begins, creating an initial identification of group

However, $\bar{p}_{-i,T}$ is a random variable determined not only by the results of earlier mixed strategy play by all other agents, but also by which agents are randomly selected to participate in each period. Therefore, it is unlikely to expect that any agent will maintain a false reputation in later periods. Knowing this, agents will have less incentive to develop reputations in early periods, unraveling the equilibrium.

membership among the subjects. A subject acting as a firm may see the group of firms as her ‘ingroup’ and the group of workers as the ‘outgroup.’ This partitioning leads naturally to categorical thinking (i.e., stereotyping) on the part of subjects, even if it is common knowledge that the outgroup is heterogeneous. As Pendry & Macrae [80, p. 926] note, “while true that outgroups are commonly perceived to be less heterogeneous in composition than ingroups, outgroup members nonetheless still display appreciable degrees of variability. Acknowledging the variability of social groups, however, is no antidote to stereotypical thinking.” Thus, firms who are aware of seller heterogeneity may not act rationally on this information.

Although experimental psychology has established that perceivers are less likely to apply existing stereotypes when the actions of the perceived affect the outcomes of the perceiver (see, for example, Neuberg & Fiske [77] or Erber & Fiske [30]), this result disappears when cognitive resources are depleted by multiple task requirements. For example, Pendry & Macrae [79] find that subjects who are first primed with a stereotype of an unknown group and then learn a long list of attributes describing various group members later recall stereotype-inconsistent attributes from the list at least as frequently as stereotype-consistent attributes. However, when subjects are also required to remember an 8-digit number while learning the list of attributes, they recall stereotype-consistent information significantly more frequently than inconsistent information. Thus, as summarized by Macrae & Bodenhausen [66, p. 105], “judgement becomes more stereotypic under cognitive load.”

Since firms in the gift exchange market are continually using cognitive resources watching the market, devising strategies, and computing payoffs, they are indeed likely to think categorically about the group of workers even though firms' payoffs depend on the behavior of workers. Furthermore, Yzerbyt *et al.* [101] find that when subjects are exposed to information about a group member inconsistent with a formed stereotype, the stereotype shifts more dramatically when the subject is under a high cognitive load. This indicates that stereotypes formed by (busy) firms are likely to change noticeably when confronted with a sudden change in behavior by a single worker.

Finally, a study by Ruscher *et al.* [88] shows that when groups are perceived to be in competition rather than individuals (e.g., when firms see themselves as *collectively* in competition with firms), then subjects tend to pay more attention to stereotype-consistent information regarding individuals in the outgroup and stereotype-inconsistent information for members of their ingroup. Additionally, Rothgerber [87] and Brewer *et al.* [10] find that "competition has the potential to create stereotypes where none or very few existed before," as summarized by Corneille & Yzerbyt [26, p. 118]. This emphasizes that there need not be existing stereotypes of the group of workers for the firms to develop stereotypes when placed in this competitive market situation.

In total, the evidence from past experiments in economics and social psychology provides reasonable support for the assumption that firms hold a collective reputation of the workers rather than tracking individual histories, and this reputation is particularly sensitive to reputation-inconsistent behaviors by individual workers. Thus, a model in which firms believe workers' types are highly correlated, though perhaps not rational, has an established psychological foundation. Taken to an extreme, the assumption of perfectly correlated types (so that any observation of $\underline{e}$ given $\overline{w}$ causes firms to believe that all agents are selfish) is in fact, sufficient to generate predictions highly consistent with the observed data in all three sessions. The following proposition formalizes this claim:.1

Proposition 4.2 *If the common knowledge prior $p_1 = \Pr[\boldsymbol{\theta}_i = \overline{\boldsymbol{\theta}}]$ is at least as large as A/B, then there exists a pure strategy reputation equilibrium of the T-period repeated mini-GEM with perfectly correlated types if and only if $C/D \leq J/I$. In this equilibrium, all firms offer $\overline{w}$ in every period, all selfish workers play $\overline{e}$ in every period $t < T$ and $\underline{e}$ in T, and all reciprocal workers play $\overline{e}$ in every period.*

As an example, let $\overline{w} = 100$, $\underline{w} = 30$, $\overline{e} = 10$, and $\underline{e} = 1$. In the HRA mini-game, $C/D \approx 0.257$ and $J/I = 2/3$. Thus, a pure strategy reputation equilibrium exists in this game. Since $A/B \approx 0.299$, if firms have at least a 3/10 prior probability that the workers are of type $\overline{\boldsymbol{\theta}}$, then cooperation should be observed by the firm in every period and by the worker in every period except the last. In the LRP mini-game, $C/D \approx 0.771$, which is greater than 2/3, so no pure strategy reputation equilibrium will exist in this mini-game. Also note that $A/B \approx 0.855$, so the restriction on beliefs would be much tighter even if J/I were greater than C/D.

In order to apply these theoretical results to the experimental data, the mini-GEM must be embedded back into the full GEM specification. While the appropriate values of $\underline{w}$ and $\underline{e}$ are clearly the stage game equilibrium values w^* and $e_{\min}$, the selection of $\overline{w}$ and $\overline{e}$ is less transparent. If the existence of the reputation equilibrium is robust to this choice for a particular specification, then the result makes more general predictions about behavior under that specification. By Proposition 4.2, the pure strategy reputation equilibrium exists for some prior p_1 if and only if $A/B < 1$ and $C/D \leq J/I$.

Figure 4.6 displays the value of $1 - A/B$ for each of the possible high wage-effort pairs $(\overline{w}, \overline{e})$ for which $A/B < 1$ and $C/D \leq 2/3$ in HRA and similarly for LRP. The value $1 - A/B$ represents the measure of the interval $[A/B, 1]$ on which prior beliefs support the reputation equilibrium using $(\overline{w}, \overline{e})$ as the 'high' strategies.[14] Larger values of $1 - A/B$ suggest that a wider range of beliefs will generate a reputation equilibrium. Reputation equilibria cannot exist for pairs $(\overline{w}, \overline{e})$ at which the graph reports a value of zero. These graphs demonstrate that the HRA specification has many more wage-effort pairs capable of supporting a pure strategy reputation equilibrium than the LRP specification, and the condition on prior beliefs is much less restrictive in HRA. Therefore, existence of reputation equilibria in the HRA are significantly more robust to perturbations of initial beliefs and the choice of $\overline{w}$ and $\overline{e}$.

In generalizing the analysis of the mini-GEM to the full GEM specification, the concept of a reciprocal worker is less concrete, given the larger strategy space. In the previous literature, a positive correlation between effort and wages is taken as an indication of reciprocity. This can be operationalized by assuming that reciprocal workers play a known pure strategy that is monotone increasing in w. For example, when $\mathcal{E} = \{1, 2, \ldots, 10\}$ and $\mathcal{W} = \{5, 10, 15, \ldots\}$, assume that reciprocal workers play

[14] Note that the set of (w, e) that Pareto dominate $(\underline{w}, \underline{e})$ is given by $\{(w, e) \in W \times E : A/B < 1 \;\&\; C/D < 1\}$. Thus, the sets where $1 - A/B > 0$ (depicted in Figure 4.6) are strict subsets of this Pareto set.

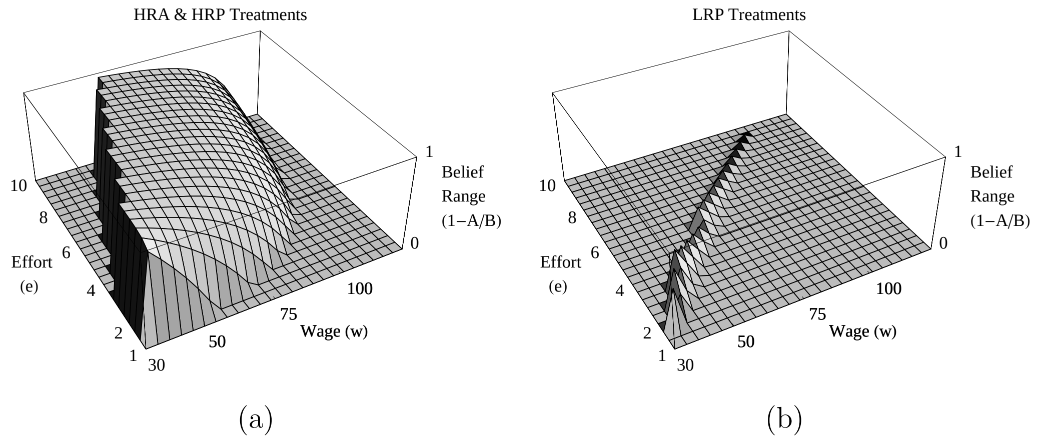


Figure 4.6: Size of the range of beliefs that support a pure-strategy reputation equilibrium in (a) the HRA and HRP treatments, and (b) the LRP treatment for various values of $\overline{w}$ and $\overline{e}$. A zero-sized belief range implies that no pure strategy reputation equilibrium exists.

a strategy weakly increasing in w given by

$$(\tilde{\chi}(w), \tilde{e}(w)) = \begin{cases} (0,1) & \text{if} \quad w < 30 \\ \left(1, \frac{w-20}{10}\right) & \text{if} \quad w \in \{30, 40, \ldots\} \\ \left(1, \frac{w-15}{10}\right) & \text{if} \quad w \in \{35, 45, \ldots\} \end{cases}$$

If a worker is known to be reciprocal, firms' profits are maximized at $\hat{w} = 75$ in HRA and at $\hat{w} = 35$ in LRP. If the firm is uncertain about the worker's type, $\hat{w}$ is weakly increasing in his belief. For example, if the firm believes the worker is reciprocal with probability 0.2, then $\hat{w}$ is 55 in HRA and 30 in LRP. In a pure strategy reputation equilibrium, the firms' beliefs do not update until after the final period, so the value of $\hat{w}$ associated with the prior belief p_1 can be sustained as a choice for $\overline{w}$ in the repeated game. Thus, $\overline{w} = \hat{w}$ is an appropriate choice for the mini-GEM analysis.[15] For this particular example, the HRA again supports a reputation equilibrium while the LRP does not.

4.4.3 Application to Previous Experiments

Since the original FKR experiment in 1993, a variety of tests of the gift exchange market have been performed by various authors. Most frequently, 'no-loss' profit functions of the form $\pi(w, e) = (v - w) e$ are used, where v is a fixed constant. This results in a situation such as that depicted in panel (a) of Figure 4.6 in which many (w, e) pairs are capable of supporting a reputation equilibrium with a wide range of beliefs. The most common result in these experiments is high wages and effort in every period, with little or no indication of reversion to stage game equilibrium (e.g., Fehr et al. [33], Charness [16], Fehr & Falk [32], Charness et al. [17], Gächter and Falk [39], and Hannan et al. [44],) indicating that most or all workers are indeed reciprocal-minded. However, the data provided by Fehr et al. [35] show strong signs of a final-period crash under the 'no-loss' payoffs. In particular, 16 out of 26 workers choose $e_{\min}$ in the final period after high wages and effort are observed in previous periods.[16] Interestingly, wages remain high in one session despite frequent observations of $e_{\min}$ throughout the game, indicating that a model of perfect correlation is in fact, not

[15] If $\tilde{e}(w)$ has a relatively steep positive slope and $\mathcal{W}$ is unbounded above, then $\hat{w}$ may diverge to infinity. In this case, it is necessary to bound $\mathcal{W}$ by some $\overline{w}$.

[16] This fact is deduced from the data in the appendix of the paper.

appropriate for this data, although a weaker degree of correlation may suffice.[17]

Several experiments have done away with the 'no-loss' condition by using linear profits of the form $\pi(w,e) = ve - w$. This does not necessarily imply that reputation equilibria are eliminated. For example, panel (a) of Figure 4.7 shows the pairs and beliefs that support reputation equilibria with the payoffs $\pi(w,e) = 10 - w + 5e$ and $u(w,e) = 10 - e + 5w$ when $\mathcal{W} = \mathcal{E} = [0,10]$, as in Brandts & Charness [9]. From Figure 4.7 it is clear that the environment supports reputation equilibria with correlation and in fact, the data show that high wages and effort move toward equilibrium on average in the final period.[18]

Riedl & Tyran [83] and Rigdon [84] also use quasilinear payoffs, and the set of wage-effort pairs sustainable as reputation equilibria with correlation is significantly smaller and larger prior beliefs are required, as demonstrated by panels (b) and (c) of Figure 4.7. In Riedl & Tyran, average wages are constant around 45 in all periods, with average efforts around 6 and several sessions featuring crashes in effort in the final period.[19] At the pair $(45, 6)$, firms' initial probability estimate that workers are reciprocal needs to be over 88 percent to support a reputation equilibrium. In Rigdon's experiment, effort decays to equilibrium early in the session, with wages following. Here, workers and firms were either unable to coordinate on a reputation equilibrium or beliefs were insufficient for such an equilibrium to exist.

[17] The fact that one worker was able to submit $e_{\min}$ in every period without destroying the group reputation indicates that it should not be an equilibrium for other selfish workers to submit higher efforts in every period, unless exactly two simultaneous occurrences of $e_{\min}$ are necessary to destroy the group reputation.

[18] Individual data is not presented, so it is unclear whether the group collectively chose slightly lower strategies or if the separation predicted by the group reputation model obtained.

[19] This fact is found in the data provided in the appendix of the paper.

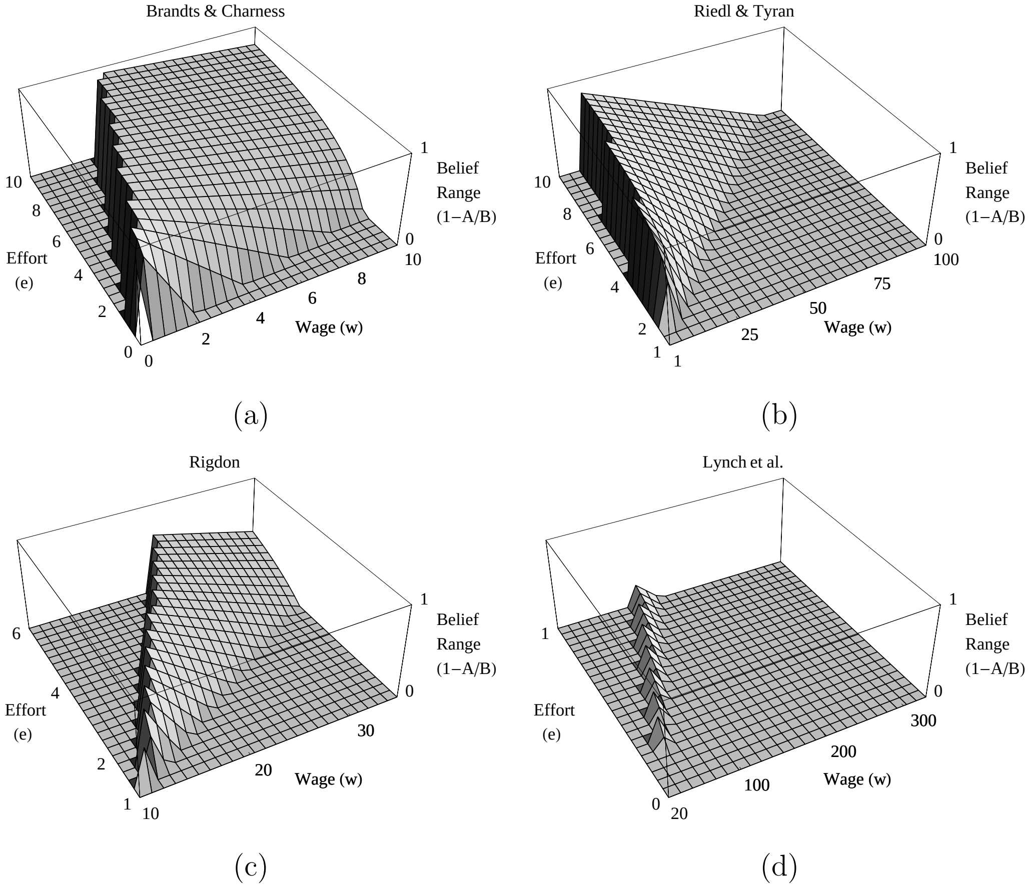


Figure 4.7: Size of the range of beliefs that support a pure-strategy reputation equilibrium in (a) the 'excess supply of labor' treatment of Brandts & Charness (2003), (b) Riedl & Tyran (2003), (c) Rigdon (2002), and (d) Lynch, Miller, Plott & Porter (2001), where 'effort' is a binary choice.

One previous experiment closely matches the conditions of the LRP environment. Lynch *et al.* [63] use a quasilinear environment with only two effort (quality) choices. Graphing the strategy pairs supporting a reputation equilibrium with correlation (panel (d) of Figure 4.7) demonstrates that at the high effort level ($\overline{e} = 1$), a reputation equilibrium can only be sustained for a very small number of wages, and only with very high prior beliefs. As predicted, wages and effort converge early to the stage game equilibrium. Lynch *et al.* conclude from their data that "a seller's demand depends not only upon his/her own 'reputation' for delivering [high quality], but also upon the market 'reputation' (p. 276)." Thus, the authors acknowledge that group reputations play an important role in these settings.

4.5 Conclusion

Although fair-minded, reciprocal behavior has been suggested as a solution to problems of contractual incompleteness, the findings of this chapter and previous experimental studies question the robustness and pervasiveness of reciprocal incentives. Inefficient equilibria are observed in some, but not all, laboratory environments.

The experimental data from this study indicate that repeated game concerns emerge in these settings despite the inability of agents to track individual reputations. A model of reputations with stereotyping recaptures the observed behavior nicely. Conditions are derived under which the problems of contractual incompleteness may be mitigated. The assumption that agents use categorical thinking (stereotyping) to assess the expected performance of others has a solid foundation in the social psychology literature. The game theoretic construct of player types provides a natural way to introduce this concept in an economic domain.

This model introduces further testable hypotheses that warrant investigation. Empirical work on consumer behavior may confirm the existence of stereotypes. For example, do customers who have had bad experiences at one auto mechanic show a reduced demand for all mechanics? Experimental studies could be used to more directly isolate the stereotype-formation phenomenon. Scoring rules could be used to elicit beliefs from subjects who purchase from a sequence of sellers under moral hazard. Functional MRI studies may provide neurological evidence for stereotype formation and its economic consequences. A variety of tests could be constructed to further examine the validity and limits of the stereotyping assumption.

Although much is to be learned about the role of categorical thinking in economic contexts, it is clear that the introduction of such a model into the environment of incomplete contracting provides more explanatory power than either a simple model of fairness or a model of individual reputation building. Thus, the 'irrational' process of stereotype formation may indeed be a powerful force by which market failure is averted.

4.6 Appendix

Assume $p_1 \geq A/B$ and let p_t be the firms' shared belief that the workers are reciprocal.

Since types are correlated, if all workers play $q_t = \Pr\left[e_{i,t} = \overline{e} | w_{j(i,t),t} = \overline{w}\right]$ in period t, then

$$p_{t+1} = \frac{p_t}{p_t + (1 - p_t)\, q_t}$$

when $e_{i,t} = \overline{e}$, which is always weakly greater than p_t. If the realization of any worker's strategy is $e_{i,t} = \underline{e}$, then $p_{t+1} = 0$. Thus, firms' beliefs weakly increase until a worker reveals that he is perfectly rational, at which point all firms know that all workers are selfish and play reverts to the fully selfish subgame perfect equilibrium.[20] In each period t, define $\mathcal{M}_t = \left\{ i \in \mathcal{I} : w_{j(i,t),t} = \overline{w} \right\}$ to be the set of workers receiving high wages in period t and note that $|\mathcal{M}_t| = J$ in every period in the proposed equilibrium.

Period T

All selfish workers play $e_{i,T} = \underline{e}$ and all reciprocal workers play $e_{i,T} = \overline{e}$ given $\overline{w}$.

If $p_T \geq A/B$, then each firm is willing to gamble on the workers by offering a high wage since

$$p_T \pi \left(\overline{w}, \overline{e} \right) + (1 - p_T)\, \pi \left(\overline{w}, \underline{e} \right) > \pi \left(\underline{w}, \underline{e} \right).$$

If $p_T = 0$, firms offer low wages.

Period $T - 1$

Selfish workers prefer to choose a mixed strategy

$$q_{T-1} = \Pr \left[e_{i,T-1} = \overline{e} | w_{j(i,T-1),T-1} = \overline{w} \right]$$

[20] Note that in the HRA specification, only one firm sees that effort choice of any given worker. However, if firm j observes $\underline{e}$ in period t, then he will offer $\underline{w}$ in period $t+1$, instantly signalling to the other firms that $p_{t+1} = 0$. Thus, as long as firms such that $p_{t+1} = 0$ offer wages first in period $t+1$, then the above result holds. If not, then other firms will update their beliefs to zero by period $t+2$. The former assumption is used here.

such that $p_T \geq A/B$ whenever the realization of all J workers' strategies in $\mathcal{M}_{T-1}$ is $e_{i,T-1} = \overline{e}$. The reader may verify that this occurs when

$$q_{T-1} \leq \left(\frac{p_{T-1}}{1 - p_{T-1}} \frac{1 - (A/B)}{A/B} \right)^{1/J}.$$

However, if $p_{T-1} \geq A/B$, then the right-hand side of this expression is weakly greater than 1, so the inequality is satisfied. Thus, regardless of the workers' strategies, firms will offer high wages in the final period whenever all workers in $\mathcal{M}_{T-1}$ provide high effort. Each worker $i \in \mathcal{M}_{T-1}$ has an expected payoff over the final two periods given by

$$\begin{aligned} & q_{i,T-1} \left[u\left(\overline{w}, \overline{e}\right) + \frac{J}{I} \left(\begin{array}{c} \prod_{i' \in \mathcal{M}_{T-1} \setminus \{i\}} q_{i',T-1} u\left(\overline{w}, \underline{e}\right) \\ + \left(1 - \prod_{i' \in \mathcal{M}_{T-1} \setminus \{i\}} q_{i',T-1}\right) u\left(\underline{w}, \underline{e}\right) \end{array} \right) \right] \\ & + \left(1 - q_{i,T-1}\right) \left[u\left(\overline{w}, \underline{e}\right) + \left(J/I\right) u\left(\underline{w}, \underline{e}\right) \right]. \end{aligned}$$

Note that this payoff is increasing in $q_{i,T-1}$ if and only if

$$u\left(\overline{w}, \overline{e}\right) + \frac{J}{I} \left(\begin{array}{c} \prod_{i' \in \mathcal{M}_{T-1} \setminus \{i\}} q_{i',T-1} u\left(\overline{w}, \underline{e}\right) \\ + \left(1 - \prod_{i' \in \mathcal{M}_{T-1} \setminus \{i\}} q_{i',T-1}\right) u\left(\underline{w}, \underline{e}\right) \end{array} \right) - u\left(\overline{w}, \underline{e}\right) - \frac{J}{I} u\left(\underline{w}, \underline{e}\right) \geq 0,$$

which is true if and only if

$$\left(\prod_{i' \in \mathcal{M}_{T-1} \setminus \{i\}} q_{i',T-1} \right) \frac{J}{I} \geq \frac{C}{D}.$$

Note that if $q_{i',T-1} = 0$ for any $i' \in \mathcal{M}_{T-1} \setminus \{i\}$, then $q_{i,T-1} = 0$ is a best response regardless of C and D since i' is fully revealing the workers' type to the firms. If $C/D > J/I$, then worker i's payoff is necessarily decreasing in $q_{i,T-1}$, so i will choose low effort with certainty. Thus, when $C/D > J/I$, $q_{i,T-1} = 0$ for all $i \in \mathcal{M}_{T-1}$ must be true in any equilibrium. If $C/D \leq J/I$, then there exists an equilibrium in which $q_{i,T-1} = 1$ for all $i \in \mathcal{M}_{T-1}$.

Assume now that $p_{T-1} \geq A/B$, $C/D \leq J/I$, and firms know all types of workers will offer high effort with probability 1. Suppose each firm j offers a high wage with probability $r_{j,T-1}$, giving firm j an expected profit over the last two periods of

$$r_{j,T-1}\pi\left(\overline{w},\overline{e}\right) + \left(1 - r_{j,T-1}\right)\pi\left(\underline{w},\underline{e}\right) + p_{T-1}\pi\left(\overline{w},\overline{e}\right) + \left(1 - p_{T-1}\right)\pi\left(\overline{w},\underline{e}\right).$$

This is strictly increasing in $r_{j,T-1}$ (regardless of the strategies of the other firms since p_T is guaranteed to equal p_{T-1}), so every firm will choose to offer high wages with probability 1.

Period $T - k$

Assume $p_{T-k} \geq A/B$ for $k > 1$. As before, regardless of strategies chosen by the workers, if the realization of all workers' strategies in $\mathcal{M}_{T-k}$ is high effort, then $p_{T-k+1} \geq A/B$, and high wages and effort will be realized in period $T - k + 1$ with probability 1. Thus, the expected payoff over the last $k + 1$ periods to a worker

$i \in \mathcal{M}_{T-k}$ when each $i' \in \mathcal{M}_{T-k}$ plays $q_{i',T-k}$ is

$$q_{i,T-k}\left[u\left(\overline{w},\overline{e}\right)+\frac{J}{I}\left(\begin{array}{c}\prod_{i'\in\mathcal{M}_{T-k}\setminus\{i\}}q_{i',T-k}\left[\left(k-1\right)u\left(\overline{w},\overline{e}\right)+u\left(\overline{w},\underline{e}\right)\right]\\+\left(1-\prod_{i'\in\mathcal{M}_{T-k}\setminus\{i\}}q_{i',T-k}\right)\left[ku\left(\underline{w},\underline{e}\right)\right]\end{array}\right)\right]$$
$$+\left(1-q_{i,T-k}\right)\left[u\left(\overline{w},\underline{e}\right)+\left(J/I\right)ku\left(\underline{w},\underline{e}\right)\right].$$

This is increasing in $q_{i,T-k}$ if and only if

$$u\left(\overline{w},\overline{e}\right)+\frac{J}{I}\left(\begin{array}{c}\prod_{i'\in\mathcal{M}_{T-k}\setminus\{i\}}q_{i',T-k}\left[\left(k-1\right)u\left(\overline{w},\overline{e}\right)+u\left(\overline{w},\underline{e}\right)\right]\\+\left(1-\prod_{i'\in\mathcal{M}_{T-k}\setminus\{i\}}q_{i',T-k}\right)\left[ku\left(\underline{w},\underline{e}\right)\right]\end{array}\right)$$
$$-u\left(\overline{w},\underline{e}\right)-\left(J/I\right)ku\left(\underline{w},\underline{e}\right)>0.$$

As in period $T-1$, this expression is positive if and only if

$$\left(\prod_{i'\in\mathcal{M}_{T-k}\setminus\{i\}}q_{i',T-k}\right)\frac{J}{I}>\frac{C}{D},$$

so when $C/D > J/I$, only $q_{i,T-k} = 0$ for all $i \in \mathcal{M}_{T-k}$ can be an equilibrium strategy. If $C/D \leq J/I$, then there exist equilibria in which $q_{i,T-k} = 1$ for all $i \in \mathcal{M}_{T-k}$.

Assume now that $p_{T-k} \geq A/B$, $C/D \leq J/I$, and firms know all types of workers will offer high effort with probability 1. Suppose each firm j offers a high wage with probability $r_{j,T-k}$, giving firm j and expected profit over the last $k+1$ periods of

$$\begin{aligned} & r_{j,T-k}\pi\left(\overline{w},\overline{e}\right)+\left(1-r_{j,T-k}\right)\pi\left(\underline{w},\underline{e}\right) \\ & +\left(k-1\right)\pi\left(\overline{w},\overline{e}\right)+p_{T-k}\pi\left(\overline{w},\overline{e}\right)+\left(\underline{e}-p_{T-k}\right)\pi\left(\overline{w},\underline{e}\right). \end{aligned}$$

This is strictly increasing in $r_{j,T-k}$ (regardless of the strategies of the other firms), so every firm will choose to offer high wages with probability 1. Thus, when $p_1 \geq A/B$ and $C/D \leq J/I$, there exists a sequential equilibrium in which high wages and high effort obtain in every period before the last. If $C/D > J/I$, then there exists no such equilibrium.