Playing 'Hard to Get': An Economic Rationale for Crowding Out of Intrinsically Motivated Behavior[☆]

Wendelin Schnedler[a], Christoph Vanberg[b,*]

[a] *University of Paderborn, Germany.*
[b] *University of Heidelberg, Germany.*

Abstract

Anecdotal, empirical, and experimental evidence suggests that offering extrinsic rewards for certain activities can reduce people's willingness to engage in those activities voluntarily. We propose a simple rationale for this 'crowding out' phenomenon, using standard economic arguments. The central idea is that the potential to earn rewards in return for an activity may create incentives to play 'hard to get' in an effort to increase those rewards. We discuss two specific contexts in which such incentives arise. In the first, refraining from the activity causes others to attach higher value to it because it becomes scarce. In the second, restraint serves to conceal the actor's intrinsic motivation. In both cases, not engaging in the activity causes others to offer larger rewards. Our theory yields the testable prediction that such effects are likely to occur when a motivated actor enjoys a sufficient degree of 'market power.'

Keywords: intrinsic motivation, crowding out, behavioral economics, market power, hidden information

JEL-Codes: D1, M5, D8, D4, C9

[☆]We gratefully acknowledge helpful comments from the editor and two anonymous referees.

[*]Corresponding author at: Department of Economics, University of Heidelberg, Bergheimer Str. 58, 69115 Heidelberg, Germany. Tel: +49 6221 542912.

Email addresses: wendelin.schnedler@upb.de (Wendelin Schnedler), vanberg@uni-hd.de (Christoph Vanberg)

1. Introduction

Consider a boy who enjoys mowing his parents' lawn and does so voluntarily. Now suppose the boy's father offers money in return for this activity. How will the boy respond to these incentives? Deci (1971) uses this anecdote in his seminal article on motivational crowding out. The term refers to the possibility that the son may become less willing to voluntarily mow the lawn after having been rewarded. Interest in (and concern about) such perverse effects of rewards has fueled both an academic and a public debate as to the underlying reasons, as well as the contexts in which they are likely to occur. A widely held view is that it may be dangerous to move activities usually engaged in 'for their own sake' into the realm of economic transactions.[1]

Crowding out effects have been demonstrated in a number of controlled experiments, starting with the seminal work of Deci (1971) and Lepper et al. (1973). As a recent example, consider an experiment conducted by Warneken and Tomasello (2008). In this experiment, young children are placed in a position where they can help an adult by picking up a fallen object. Previous studies have shown that most children are *intrinsically* motivated to engage in such helping behavior, meaning that they will do so spontaneously and in the absence of any promise of being rewarded (Warneken and Tomasello, 2006). Warneken and Tomasello (2008) randomly assign some children to a treatment condition in which they are explicitly offered a material reward in return for helping. In a subsequent phase where no such rewards are offered, these children are found to be less likely to help than those in a no-reward control group. Fabes et al. (1989) conduct a study in which children could help by sorting pieces of paper according to color. In one condition, the experimenter explicitly offered a reward in exchange. In a subsequent phase, the experimenter left the room, announcing that he will return later, and the children were secretly observed. Compared to children who had not seen rewards being offered, these children were less likely to sort papers in the absence of the adult.

Effects similar to those documented in these studies have been found in a number of psychological and economic experiments. See Deci et al. (1999) for a meta study on psychological experiments, Bowles and Polania-Reyes (2012) for a survey on economic experiments, and Gächter et al. (2011) for a recent

[1]This view is, for example, advocated by Alfie Kohn in his bestsellers 'Punished by Rewards' (1999) and 'Unconditional Parenting' (2005).

example of an economic experiment on the dynamic effects of incentives on voluntary cooperation.[2]

A somewhat ad-hoc explanation for such effects would be that economic incentives cause a change in preferences. Under normal conditions, children experience an internal ('intrinsic') reward when they engage in pro-social activities such as picking up dropped objects or sorting papers. However, if another person explicitly offers a reward in return for such action, this reduces or eliminates the internal reward. Indeed, Frey (1994) argues that crowding out would be 'difficult or impossible to account for in a reasonable way' without assuming such a change in fundamental preferences.[3] A disadvantage of this ad hoc theory is that it is silent as to the ultimate *reasons* for such changes, and sheds no light on the conditions under which they are likely to occur.

A number of authors have proposed explanations for crowding out that do not involve a change in fundamental preferences. One prominent explanation is that individuals use activities as signals in order to create or maintain a positive (self) image. If paid for, these activities may lose their signaling value (Seabright, 2004, 2009; Bénabou and Tirole, 2006, 2011). This may, for example, explain why someone might be less willing to donate blood when offered money in return. However, it does not seem to apply to the experiments mentioned above. If the child helps in order to signal that it is 'good,' then its ability to do so is compromised only when a reward is offered. Once rewards are removed, the activity can be used as a signal again and the child should engage in it. In the experiment, however, children are less likely to engage in the activity after rewards are removed.

Several explanations are based on the idea that payment may constitute a signal from a better informed party (e.g. parent, teacher or employer) that affects the beliefs of the actor (e.g. child, pupil, worker) concerning the nature of the task. For example, payments may indicate that the activity is dangerous or otherwise costly (Bénabou and Tirole, 2003).[4] However, this

[2]A closely related literature documents possibly deleterious effects of monitoring and control (Falk and Kosfeld, 2006; Schnedler and Vadovic, 2011; Ploner et al., 2011). Von Siemens (2013) attributes such effects to negative reciprocity.

[3]Sliwka (2007) proposes a model in which 'conformists' infer from incentives that others are not intrinsically motivated. By assumption, this causes them to become unmotivated.

[4]In Herold (2010), payment indicates that the informed party expects the agent to fail. In Friebel and Schnedler (2011), it suggests that colleagues are not 'team players.' In

information based explanation for crowding out seems unlikely to apply to experiments such as those we have mentioned. While it is possible that payment reveals to the child that picking up objects or sorting papers is difficult, onerous or dangerous, this appears rather implausible. After all, the activity is extremely simple, and the child has probably experienced it before. Indeed, Bowles and Polania-Reyes (2012) find that theories of information revelation are only consistent with about a third of the economic experiments which they review in their survey article.[5]

According to Bowles and Polania-Reyes (2012), much of the experimental evidence on motivational crowding out seems to be driven by what they call 'framing': economic incentives establish a 'market frame' and induce a 'market mentality.' Even though the fundamental nature of the activity is unaffected, the frame 'activates own payoff-maximizing modes of thought.' This view differs from the idea that payments modify fundamental preferences in that the individual is seen as having a 'repertoire of preferences.' Rather than literally altering these preferences, incentives affect their 'salience' within a specific situation. Just like its more ad-hoc cousin, however, this theory begs the question: Why should a mere 'market frame' undermine (or render less salient) an individual's pro-social motivations?

The present paper proposes a simple and intuitive economic rationale for why a reward-induced 'market frame' may lead to crowding out effects. This rationale does not appeal to changes in the (actual or perceived) attractiveness of the activity *per se*. Instead, our theory is based on the idea that the 'market frame' triggers certain beliefs about the rules governing an interaction. In particular, we assume that the 'frame' determines whether or not individuals consider it possible or appropriate, in principle, that the activity in question might be paid for.[6]

The essence of our argument is the following. Consider an individual who

---

Schnedler and Vadovic (2011), it indicates that the informed party does not expect the agent to engage in it. van der Weele (2012) argues that payment may suggest that the activity is not a prevailing norm.

[5] A further explanation for crowding out posits that greater wealth renders an activity less attractive (Schnedler, 2011). As the child accumulates no wealth, this explanation can also be ruled out.

[6] Bowles and Polania-Reyes (2012) suggest that frames provide "cues for appropriate behavior." In our theory the "cue" does not relate to the activity directly but to the question of whether others may, in principle, offer rewards in exchange. The resulting effects on behavior are rational, strategic responses to the perceived rule change.

attaches *intrinsic value* to the performance of certain activities, such as helping an adult. In absence of other ('ulterior') considerations, the individual is motivated to engage in the activity *per se*, and would do so voluntarily. If this activity becomes the object of economic transactions, however, the individual will, in addition to its *intrinsic value*, consider its *exchange value* (its 'price'), and he will normally wish to increase this exchange value if he can. Under certain conditions, this goal will be served by 'cutting back' on the extent to which he engages in the activity. Thus, *even if the actor's intrinsic motivation is unchanged relative to the situation without extrinsic rewards*, he may be less willing to perform the action in their presence. We describe this strategic response as 'playing hard to get' and identify two contexts in which there is an incentive to do so.

Both contexts are discussed in terms familiar to economists, highlighting the strictly economic logic of our explanation. In keeping with this perspective, we will refer to the individual engaging in the activity of interest as a 'producer' (he), to the person benefiting from it as a 'consumer' (she), and to the exchange value as its 'price.' In the experiments mentioned above, the child is in the role of the producer, the adult the consumer, and the price corresponds to the reward being offered. In both contexts, we compare what happens in two possible 'frames'—one in which the activity is an object of exchange with a 'price,' and one in which it is not. In the latter case, we say that the activity is not 'tradeable.' These 'frames' correspond to the different treatments experienced by children in the experiments. We say that extrinsic rewards *crowd out* intrinsically motivated behavior if an intrinsically motivated producer's 'supply' of the activity is lower when it is tradeable as compared to when it is not. In both contexts, we show that an intrinsically motivated producer may wish to reduce his 'supply' (thus foregoing some intrinsic reward) in order to increase the activity's 'price' (extrinsic rewards). The two contexts differ in terms of *why* 'playing hard to get' has this effect on prices.

In the first context, 'playing hard to get' causes the exchange value to rise because the activity becomes scarce. As an example, consider the experiment by Fabes et al. (1989). During the no-reward phase, it is reasonable for the child to expect that the adult, upon returning, will place more value on sorting of additional papers if fewer papers are sorted ('for free') in the interim. In the reward condition, the child may then reasonably expect that the adult will be more likely to offer a reward for additional sorting if she does not sort (or sorts less) in the interim. The observed reluctance to sort

can thus be understood as strategic exploitation of a monopoly position. In Section 2, we introduce intrinsically motivated producers to the textbook monopoly model and show that crowding out occurs whenever demand is inelastic (Proposition 1).

In our second context, playing 'hard to get' increases the activity's exchange value because it suggests that the producer finds the activity unpleasant. To see how this works, consider the experiment by Warneken and Tomasello (2008). Here, our argument suggests that children refrain from helping in order to pretend that they do not like it. Although they would want to help, they can hope to convince the adult that compensation is required if she wants them to pick up the object. In Section 3, we illustrate this idea in a model involving one consumer and a single producer whose motivation is private knowledge. We show that crowding out increases with the consumer's benefit from the activity (Proposition 2).

In both contexts, a 'playing hard to get effect' emerges only if the intrinsically motivated individual is hard to replace, meaning that the 'service' he is motivated to provide cannot simply be offered by others. We regard this as a substantively interesting and empirically testable implication of our analysis. Difficulties in substituting pro-social activities of individuals arise particularly in personal relationships. Our analysis thus provides a formal underpinning for Frey's proposition (1994) that such activities are particularly vulnerable to crowding out.

It is worth stating at the outset that our theory does not aim to explain why or when a particular activity is regarded as 'tradeable.' In the experiments mentioned, this is a function of the treatment a subject is assigned to. Children assigned to the 'reward' treatments experience a 'different world' (or *frame*) than those in the control treatment. In this 'market frame', adults explicitly offer material rewards in exchange for behavior that children would otherwise engage in spontaneously. In our effort to explain the differences in behavior between these conditions, we take the non-tradeable or tradeable nature of an activity as given and explore its consequences. The analysis does, however, allow us to derive conclusions as to whether tradeability is desirable. While the producer benefits from tradeability, consumers are harmed due to the reduction in supply as well as the payments that become necessary. Overall, the establishment of a 'market' for a pro-social activity can be socially harmful in the sense that tradeability leads to a lower aggregate surplus (Corollaries 1 and 2).

A common theme in prior explanations for crowding out effects has been

that rewards effectively 'spoil the fun' that an individual derives (or expects to derive) from the activity. For example, rewards indicate that the activity is costly, or they destroy its signaling value. In both cases, rewards remove the individual's original motive for engaging in the activity. Our explanation differs from these in that rewards do not affect this original motive. Instead, tradeability introduces an additional and competing *exchange* motive, as the activity now also affects the size of extrinsic rewards.

We regard this exchange motive as a potentially important source of crowding out effects that is distinct from, but not incompatible with previous theories mentioned above. Complementing the literature, it may explain crowding out in contexts where rewards are unlikely to affect the 'innate' attractiveness of an activity, as when Bowles and Polania-Reyes (2012) attribute crowding out to 'market frames.' It also yields a new and testable prediction: in situations where our argument applies, crowding out should be less likely if there is competition, as playing 'hard to get' would induce consumers to turn to a competing substitute. Our analysis is also the first to explicitly link crowding out to negative welfare effects.[7]

The remainder of the paper is organized as follows. Section 2 presents a model in which a producer plays 'hard to get' to keep his activity scarce. In Section 3, we formalize the idea that the producer plays 'hard to get' to hide his intrinsic motivation. Section 4 concludes. Formal proofs are contained in the Appendix.

2. Creating Scarcity

One way in which refraining from an activity may increase its exchange value is that others attach more value to it if it becomes 'scarce.' If a child in Fabes et al. (1989) sorts lots of papers during the experimenter's absence, the adult may be less likely to reward when returning than if the child sorts few or no papers. If so, we can say that the child has some degree of 'market

[7]Our paper also adds to the literature on withholding effort in dynamic settings. Weitzman (1980) postulates that firms adapt the threshold for bonus payment to past performance and shows that workers then have an incentive to withhold effort; the famous 'ratchet effect.' Suvorov (2003) presents a model in which agents withhold effort to conceal confidence. In Angerhausen et al. (2010), unemployed workers reject job offers to signal their value to prospective employers. In contrast to all these contributions, intrinsic motivation is key to our argument, as future rewards are reaped by foregoing an intrinsically pleasurable activity.

power'—resulting from the fact that it alone is in a position to sort the papers in question. Thus, there is an incentive to play 'hard to get' (by not sorting during the experimenter's absence) in order to 'drive up the price' for paper sorting. In this section, we formalize this argument in the simplest possible way: by introducing intrinsic motivation into the textbook model of a (non-discriminating) monopolist.

A producer (he) supplies y. The supply could be the number of sorted colored papers 'produced' by the child. As in the textbook monopoly model, we assume that the producer's utility is given by $u(r, y, \theta) = r(y) - c_\theta(y)$, where $r(y)$ denotes (extrinsic) revenue collected, and $c_\theta(y)$ denotes (intrinsic) costs incurred depending on the degree of intrinsic motivation θ. Unlike the textbook model, we assume that he intrinsically *benefits* from supplying a certain amount of the activity, even in the absence of payment. The child is, for example, willing to voluntarily sort a certain number of papers. Formally, we assume that marginal costs are initially *negative* and increasing, so that total costs are minimized at some positive level of supply, which we denote by y^ν.

The consumer (she) has an inverse demand function $p(y)$, representing her willingness to pay. Let the producer's revenue amount to $r(y) = p(y) \cdot y$ if y is tradeable and to zero, otherwise. Suppose that the marginal revenue, $r'(y)$, is falling. The child may know (or perceive) that the adult is less willing (or less likely) to reward if he has already sorted a lot of papers. Formally, this assumption ensures that the producer's 'profit function' is concave and has an inner solution in case that y is tradeable.

We now examine what would occur in two possible 'frames.' In the first frame, the activity is not tradeable. Helping is not considered an economic activity. It may for example be 'taboo' or simply 'unheard of' that children are given material rewards in exchange for helping. In the experiment, this is the frame experienced by children in the baseline condition. In the second frame, the activity is tradeable, meaning that it is normal to pay for help. Children assigned to the experimental 'reward' condition are exposed to this frame. In line with the experimental study, we will compare these frames and say that 'crowding out' occurs if the producer engages in the activity less often when it is tradeable.

Take first the frame in which helping is not an economic activity and has no exchange value. In this case, the producer's behavior is driven entirely by his intrinsic cost function, and he chooses $y^\nu > 0$. For illustrative purposes, let demand be linear $p(y) = 1 - y$ and costs quadratic $c_\theta(y) = (y - \theta)^2$,

where $\theta > 0$ reflects that the producer intrinsic motivation. The situation is illustrated in Figure 1. The child chooses his desired level of helping y^ν, at which point he is satisfied (his marginal costs are zero): $c'(y^\nu) = 0$.

However, y^ν is smaller than the Pareto optimal level y^o, where marginal cost and inverse demand curve intersect: $c'(y^o) = p(y^o)$. The reason for this inefficiency is, of course, that the child's activity has a positive externality. This diagnosis might suggest, at first sight, that the creation of a *market* for helping improves efficiency.

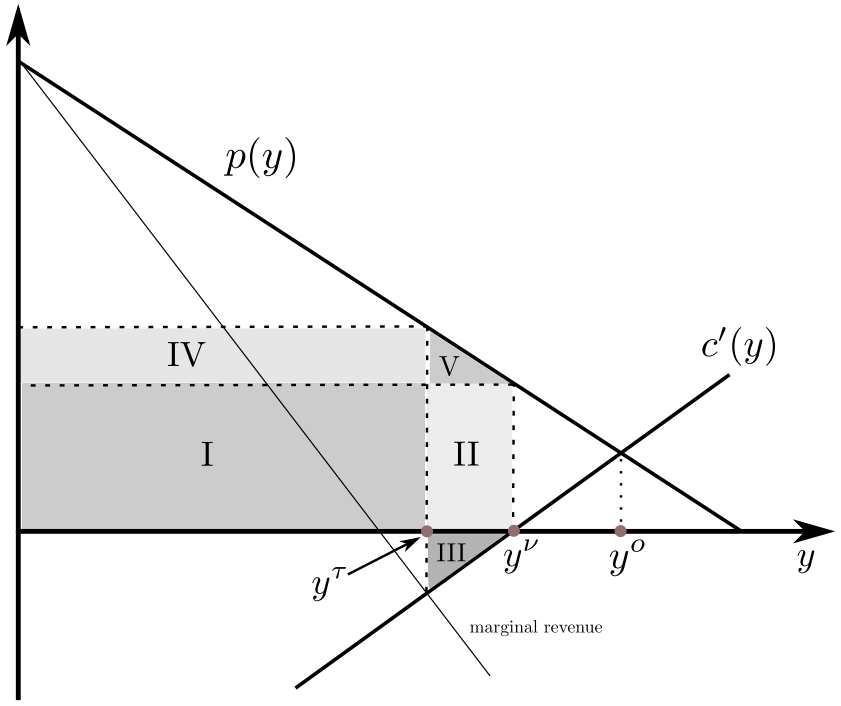

Figure 1: Example of a monopolist who reduces supply to create scarcity after his service becomes tradeable.

Next, consider the frame where the activity is tradeable. The producer will then consider the effects of his behavior on the price that he can command in the market for the activity. Being a (local) monopolist, he chooses the activity level at which marginal cost equals marginal revenue. In our example, this is $y^\tau = \frac{\theta+0.5}{2}$. Here the superscript τ is used to denote the level chosen under tradeability, i.e., when both intrinsic rewards and economic incentives are present. Consequently, tradeability leads to a change in the

supply that amounts to

$$y^\tau - y^\nu = \frac{\theta + 0.5}{2} - \theta = \frac{1}{2}(0.5 - \theta).$$

This implies that a sufficiently motivated producer ($\theta > 0.5$.) supplies less whenever the activity is tradeable.

Before generalizing this argument, let us examine how tradeability affects producer and consumer welfare. The producer is unambiguously better off when the good is tradeable than when it is not. To see this, imagine for a moment that the producer offers y^ν even when the good is tradeable. Then, the only difference to the non-tradeable situation would be that he captures some of the consumer surplus (areas I and II in Figure 1), making him better off. By optimally reducing supply to y^τ, he captures additional extrinsic rewards (area IV minus area II), the value of which must exceed the intrinsic value of the foregone activity (area III), making him better off again. Consumers on the other hand are worse off if the activity is tradeable, as supply is lower *and* it is no longer free. Their loss is represented by areas I, II, IV, and V. Overall, society incurs a welfare loss from the reduced supply, captured by areas II, III, and V. In fact, both the producer and consumers would benefit if the producer simply provided some additional units of the activity for free. The additional units, however, cannot be provided without reducing the exchange value, which is why the producer abstains from producing them.

The substantive conclusions derived in the linear demand example extend to the more general model. If the activity is not tradeable, the producer chooses the cost minimizing level y^ν. Since y^ν minimizes costs (or equivalently maximizes the producer's pleasure), the direct effect from marginally increasing or decreasing supply on the utility at the voluntary choice y^ν is zero. If the activity is tradeable, the producer's decision to supply either more or less than y^ν is thus entirely driven by the effect of such a change on his revenue. The producer reduces supply if and only if marginal revenue is negative at y^ν. Equivalently, the reduction is optimal for the producer when the consumer's demand is inelastic. This establishes the following result. (Proofs of all propositions can be found in the appendix.)

Proposition 1 (Crowding Out)**.** *If an activity becomes tradeable, a motivated producer reduces supply whenever demand at the voluntary activity level is inelastic:*

$$y^\nu > y^\tau \text{ if and only if } |\epsilon| < 1,$$

where $\epsilon = \frac{p(y^\nu)}{p'(y^\nu)y^\nu}$ *is the elasticity of demand at* y^ν*.*

To interpret this result, note that the elasticity of demand will depend on the consumer's ability to find a suitable substitute. The more difficult this is, the smaller is ϵ in absolute value. One interpretation of the proposition is hence that crowding out is more likely to occur if a producer is hard to substitute. For example, it is more likely to occur if only one child is available to sort papers than if several children were eager to perform this task.

Next, consider the welfare implications of crowding out. As in the example, it is clear that crowding out implies that tradeability makes consumers worse off, because they have to pay for the activity *and* obtain less of it. On the other hand, the producer is unambiguously better off if the activity is tradeable. To see this, recall that he is free to engage in the voluntary activity level y^ν even in the market scenario, and he would then be better off due to the fact that he receives payment. If he finds it optimal to reduce supply, his utility increases further. While consumers lose and producers benefit, the net effect on aggregate surplus is negative because both the producer and consumers would benefit if additional units were provided for free. Let us summarize these considerations.

Corollary 1 (Welfare consequences)**.** *If crowding out occurs, tradeability (i) reduces consumers' surplus, (ii) increases producer's surplus and (iii) leads to a loss of aggregate surplus.*

The basic conclusions from this model extend to other market types, as long as the producer has sufficient market power (Cournot-oligopoly, monopolistic competition, etc.). Substantively, the incentive to 'play hard to get' arises in contexts where those who benefit from an activity cannot easily turn elsewhere. For example, one might expect this type of crowding out effect to occur in the context of personal relationships, where the intrinsically motivated individual is in some sense unique or special.[8]

3. Hiding Motivation

There is a second way in which a reduction in supply of an activity may increase its exchange value. By refraining from an activity from which others

[8]In personal relationships, it is particularly relevant that everybody is a 'monopoly supplier of his own actions'— an expression that we owe to Dan Houser.

benefit, an individual may create the impression that he does not regard the activity as intrinsically rewarding, perhaps causing others to offer him extrinsic rewards. If a child in the experiment by Warneken and Tomasello (2008) refuses to pick up a fallen object, the adult may conclude that the child dislikes helping and must be compensated to do so. As this may cause the adult to offer compensation, the child has an incentive not to pick up the object, even if it intrinsically likes to help. If successful, it will obtain (perhaps after some time) a monetary reward, in addition to the intrinsic reward from helping. In this section, we illustrate this mechanism in a simple model.

A producer (in our example: the child, he) and a consumer (the adult, she) interact for two periods $t = 1, 2$. In each period, the producer chooses whether or not to perform a task (e.g. picking up an object). We denote his choice by $y_t \in \{0, 1\}$. If the producer performs the task ($y_t = 1$), the consumer experiences a benefit $B \geq 0$. Substantively, B represents the *net gain* to the consumer from having the task completed by the producer in question (e.g. the child), instead of moving to the next best alternative (e.g. walking to and picking up the object herself). We can interpret B as a measure of 'non-substitutability' of the producer's action.[9]

If he performs the task, the producer receives an intrinsic payoff denoted by θ. We assume that the value of θ is ex ante uniformly distributed on $[-1, 1]$. The producer can hence be motivated and benefit from the task, $\theta \in (0, 1]$, or unmotivated and suffer from the task, $\theta \in [-1, 0)$. The producer's motivation is only known to the producer himself and not to the consumer. In addition to the intrinsic payoffs, producers derive utility from (extrinsic) rewards, denoted r_t. The producer's period t utility is $u_t(y_t, r_t) = \theta \cdot y_t + r_t$, and the consumer's period t utility is $v_t(y_t, r_t) = B \cdot y_t - r_t$.

As above, we consider two 'frames,' one in which the activity is not tradeable, meaning that the consumer cannot offer payment in exchange, and another in which the activity is tradeable. In the latter case, we assume for simplicity that the consumer may only offer rewards in the second period.[10]

[9] If $B = 0$, the producer is perfectly substitutable, because the next best alternative yields the same benefit to the consumer. If $B > 0$, the producer is 'non-substitutable' in the sense that he can generate a benefit that the next best alternative cannot.

[10] In our working paper, we present a model in which the consumer can offer a reward in both periods. All results derived here extend to that model as long as B is not too large. For sufficiently large values of B, the consumer will offer rewards great enough

In both cases, we derive the Perfect Bayesian Nash equilibria of the game, which turn out to be unique in terms of the relevant equilibrium behavior.[11] We say that crowding out occurs if motivated producers (i.e., some set with $\theta > 0$ of positive measure) engage in the activity less often when it is tradeable.

We begin by considering the case in which the task is not tradeable. Clearly, an intrinsically motivated producer then performs the activity in both periods: $y_1^\nu(\theta) = y_2^\nu(\theta) = 1$ if $\theta > 0$. On the other hand, an unmotivated producer ($\theta < 0$) refrains from doing so: $y_1^\nu(\theta) = y_2^\nu(\theta) = 0$ if $\theta < 0$. (Producers of type $\theta = 0$ may take either action in both periods.) If $B > 0$, this outcome is not Pareto efficient. Efficiency would require that unmotivated producers with $\theta \in (-B, 0)$ perform the task. Their costs are below the consumer's benefit, so that they could be compensated by the consumer. As in the previous setting, one might suspect that introducing the possibility of payments improves the situation.

Next, we consider the case in which the producer's activity is tradeable. The consumer can thus offer a price, $p \geq 0$ in the second period. Intuition suggests that a producer who engages in the activity in the first period thereby reveals his intrinsic motivation, leading the consumer to offer no reward in the second period. By contrast, a producer who refrains from the activity will appear intrinsically unmotivated, perhaps inducing the consumer to offer an extrinsic reward. If so, even an intrinsically motivated producer may wish to refrain from the activity in the first period. By playing 'hard to get,' he can hide his motivation and elicit payment in the second period. The following result shows that this intuition is correct.

Lemma 1 (Behavior given tradeability)**.** *If the activity is tradeable, there exists a Perfect Bayesian equilibrium. Further, the following statements hold in any Perfect Bayesian Equilibrium.*

(i) In period 2, producers of type $\theta > -p$ perform the activity while those of type $\theta < -p$ do not, where p is the price offered by the consumer. (Producers of type $\theta = -p$ may choose either action.)

to counteract the 'playing hard to get' effect. Off the equilibrium path, however, some intrinsically motivated producers would not perform the task if the reward was not offered.

[11] The equilibria are unique up to different rules on how to break indifference. Which rule is used has no impact on other properties of the equilibria.

(ii) In period 2, the consumer offers a reward equal to $p = \min\{\frac{B}{3}, 1\}$ *if and only if the producer has not performed the activity in period 1.*

(iii) In period 1, producers of type $\theta < \frac{B}{3}$ *do not perform the activity, while those of type* $\theta > \frac{B}{3}$ *do. (Producers of type* $\theta = \frac{B}{3}$ *may chose either action.)*

The effects of moving the activity into the 'market domain' can be directly assessed by comparing the behavior when the activity is tradeable (as described in Lemma 1) with that in absence of tradeability. This immediately yields the following proposition.

Proposition 2 (Crowding Out)**.** *When an activity becomes tradeable,*

(i) fewer motivated producers perform the activity in period 1:

$$\text{for } \theta \in \left(0, \frac{B}{3}\right) : y_1^{\nu}(\theta) = 1 \text{ and } y_1^{\tau}(\theta) = 0,$$

(ii) more unmotivated producers perform the activity in period 2:

$$\text{for } \theta \in \left(-\frac{B}{3}, 0\right) : y_2^{\nu}(\theta) = 0 \text{ and } y_2^{\tau}(\theta) = 1,$$

(iii.) the consumer pays for the activity in period 2 whenever $\theta \in (-\frac{B}{3}, \frac{B}{3})$.

This result parallels Proposition 1 in three respects. First, tradeability leads to 'crowding out' of voluntary provision by intrinsically motivated producers. Second, crowding out only occurs if the producer generates a (relationship-specific) benefit $B > 0$. This benefit implicitly reflects the extent to which the producer's service is difficult to substitute, as it is bounded by the benefit that can be derived from the next best alternative. Third, the more difficult it becomes to substitute the producer, the larger is the crowding out effect. As in the previous situation, the negative effect concerns motivated producers. Unmotivated producers, on the other hand, increase their supply.[12]

[12]In the linear demand model, tradeability leads to a reduction in supply only when the producer is sufficiently motivated in the sense that $\theta > 0.5$. Less motivated producers increase their supply.

What are the effects of tradeability on producer and consumer utility? Producers who strongly dislike the activity ($\theta < -B/3$) are unaffected as they do not engage in the activity whether it is tradeable or not. Producers with a strong intrinsic motivation ($\theta > B/3$) are also unaffected, as they perform the task in both periods and receive no extrinsic rewards in either case. Producers who 'weakly dislike' the activity ($\theta \in (-B/3, 0)$) strictly benefit from tradeability, as they now engage in an intrinsically costly activity in period 2, but receive a payment in excess of their costs. Finally, consider producers with some intrinsic motivation $\theta \in (0, B/3)$. These are the individuals whose period 1 activity is 'crowded out.' Accordingly, they forgo the intrinsic pleasure of the activity in period 1. However, they do so because they value the second period payment more highly than the pleasure foregone. Thus, some producer types are made strictly better off by tradeability, and none are made worse off.

To assess the effect of tradeability on the consumer's welfare, notice that her expected benefit from the producer's activity is unaffected: The interval of motivated types who refrain from performing in period 1 is exactly as large as the interval of unmotivated types who are induced to perform in period 2. Thus the expected sum of period 1 and period 2 activity is unchanged by tradeability. On the other hand, the consumer now sometimes has to pay for the activity. In expectation, she is strictly worse off.

Finally, consider the effect of tradeability on aggregate surplus. Since payments constitute mere transfers, we need only consider the benefits and costs generated by the activity itself. As we have seen, the expected consumer benefits are unchanged, since the expected 'amount' of activity is unaffected by tradeability. On the producer side, however, tradeability means that the activity is sometimes provided by *unmotivated* types (in period 2) instead of by motivated types (in period 1). It follows that the expected sum of payoffs falls relative to the non-tradeable environment. We summarize these welfare results in the following corollary.

Corollary 2 (Welfare consequences)**.** *In expectation, tradeability (i) strictly increases the producer's utility, (ii) strictly diminishes the consumer's utility, and (iii) strictly lowers aggregate surplus.*

4. Conclusion

Previous research in psychology and economics, as well as public discussion, has been concerned about the possibly detrimental effects of introducing

extrinsic rewards into settings where people engage in socially desirable activities 'for their own sake.' The concern is that turning such activities into objects of 'economic exchange' may backfire and actually cause people to stop engaging in them. Anecdotal and experimental evidence suggests that this 'crowding out' phenomenon can occur in practice. The reasons underlying such effects and the conditions under which they are likely to arise, however, remain the subject of an ongoing debate.

Detrimental effects of rewards have been explained in a number of ways. Although not all previous explanations make use of an ad hoc preference change, a common theme is that rewards 'spoil the fun' of the activity in the sense that the individual expects to derive a lower intrinsic benefit from engaging in it. In contrast to these explanations, we have shown that rewards may 'crowd out' intrinsically motivated behavior even if people derive the same benefit from the activity, i.e., even if they continue to 'like' doing it.

In a nutshell, the explanation for crowding out we offer here is that the prospect of earning extrinsic rewards creates incentives to play 'hard to get' in order to increase the exchange value of an activity. Doing so may make the activity scarce and hence more valuable to others. It may also conceal one's own motivation and thereby lead others to offer larger rewards. In both cases, the strategy involves inefficient behavioral adjustments, as intrinsically motivated agents refrain from engaging in an enjoyable activity which would also generate benefits to others. This negative effect may be partially offset if extrinsic rewards induce intrinsically *un*motivated people to perform the task. However, if a sufficient number of people are intrinsically motivated, the introduction of extrinsic rewards may be harmful in an aggregate surplus sense.

To conclude, let us briefly review how our analysis applies to the experiments described in the introduction. In these experiments, children assigned to the reward condition experience an environment within which adults can and do offer extrinsic rewards in exchange for help. Our analysis shows that the observed reduction in helping behavior during a subsequent no-reward condition may be a *strategic response* designed to bring back (or increase) such rewards. Given what the children know about the situation they are in, this response is entirely reasonable. During the 'no reward' phase of the paper sorting experiment, the child has been told that the adult will return, and thus it can reasonably expect to be offered a reward for additional sorting at that time. In the experiment of Warneken and Tomasello (2008), children who are offered a reward for helping at the beginning do not know that the

adult will not offer rewards later. As far as they are concerned, it is reasonable to expect that refusing to help 'for free' may induce the adult to offer a reward again.

In the context of these and similar experiments, we feel that our theory provides a more plausible explanation for the behavior observed than do previous theories. Consider, for example, the idea that rewards undermine the (self) signaling value of pro-social behaviors. This theory predicts reduced helping while rewards are offered. Within the experiments, however, the reduction in helping is observed in the subsequent no-reward phases, at which point the signaling value of helping is restored. Next consider the idea that rewards reveal the task involved to be more costly (or less enjoyable) than initially believed. While this theory is consistent with the timing of the observed effects, it appears to us rather implausible. After all, the tasks involved are extremely simple, and the children have already experienced them before. Previous explanations of crowding out may well apply in other contexts. Within the (very typical) experiments we have mentioned, it seems to us more plausible that subjects are 'playing hard to get' in an effort to reap further and larger gains from helping behavior that they would otherwise have engaged in anyway.

Strictly interpreted, our theory does not apply to experiments in which subjects are certain that their behavior cannot affect rewards. For example, instructions may explicitly state that the experiment will end after the no reward phase.[13] We believe, however, that our explanation may also shed light on behavior under these circumstances. It is likely that experimental subjects have experienced real world situations to which our theory applies. Based on their experiences, they may have developed a playing-hard-to-get *heuristic*, which they mistakenly apply in the lab.[14]

[13]This is not true for field experiments such as the helping experiments we have mentioned. The same applies to most of the psychological evidence we are aware of, as well as economic experiments with an unknown number of parts (see e.g. Gächter et al., 2011).

[14]Indeed, our theory could be interpreted as providing a strategic rationale for the existence of such a heuristic or why evolution shaped humans to 'lose interest' in activities which are used in social exchange.

References

Angerhausen, J., Bayer, C., Hehenkamp, B., 2010. Strategic unemployment. Journal of Institutional and Theoretical Economics 166, 439–461.

Bénabou, R., Tirole, J., 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70, 489–520.

Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. American Economic Review 96, 1652–1678.

Bénabou, R., Tirole, J., 2011. Identity, morals, and taboos: Beliefs as assets. The Quarterly Journal of Economics 126, 805–855. URL:

Bowles, S., Polania-Reyes, S., 2012. Economic incentives and social preferences: Substitutes or complements. Journal of Economic Literature 50, 368–425.

Deci, E.L., 1971. Effects of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology 18, 105 – 115.

Deci, E.L., Koestner, R., Ryan, R.M., 1999. A meta-analytic review of experiments examing the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627–668.

Fabes, R.A., Fultz, J., Eisenberg, N., May-Plumlee, T., Christopher, F.S., 1989. Effects of rewards on children's prosocial motivation: A socialization study. Developmental psychology 25, 509.

Falk, A., Kosfeld, M., 2006. The hidden costs of control. American Economic Review 96 (5), 1611–1630.

Frey, B.S., 1994. How intrinsic motivation is crowded out and in. Rationality and Society 6, 334–352.

Friebel, G., Schnedler, W., 2011. Team governance: Empowerment or hierarchical control. Journal of Economic Behavior and Organization 78, 1–13.

Gächter, S., Kessler, E., Königstein, M., 2011. The roles of incentives and voluntary cooperation for contractual compliance. Discussion Paper 06. University of Nottingham, Centre for Decision Research and Experimental Economics.

Herold, F., 2010. Contractual incompleteness as a signal of trust. Games and Economic Behavior 68, 180–191.

Kohn, A., 1999. Punished by rewards: The trouble with gold stars, incentive plans, A's, praise, and other bribes. Houghton Mifflin Co.

Kohn, A., 2005. Unconditional parenting: Moving from rewards and punishments to love and reason. Atria Books.

Lepper, M.R., Greene, D., Nisbett, R.E., 1973. Undermining children's intrinsic interest with extrinsic rewards: A test of the "overjustification" hypothesis. Journal of Personality and Social Psychology 28, 129–137.

Ploner, M., Schmelz, K., Ziegelmeyer, A., 2011. Hidden costs of control: Three repetitions and an extension. Experimental Economics 15 (2), 323–340.

Schnedler, W., 2011. You don't always get what you pay for. German Economic Review 12, 1–10.

Schnedler, W., Vadovic, R., 2011. Legitimacy of control. Journal of Economics and Management Strategy 20, 985–1009.

Seabright, P.B., 2004. Continuous preferences can cause discontinuous choices: An application to the impact of incentives on altruism. CEPR Discussion Papers 4322. C.E.P.R. Discussion Papers.

Seabright, P.B., 2009. Continuous preferences and discontinuous choices: How altruists respond to incentives. The B.E. Journal of Theoretical Economics 9, Article 14. Contributions.

Sliwka, D., 2007. On the hidden costs of incentive schemes. American Economic Review 97, 999–1012.

Suvorov, A., 2003. Addiction to rewards. Toulouse School of Economics. mimeo.

Titmuss, R.M., 1971. The gift relationship: From human blood to social policy. Pantheon Books, New York.

van der Weele, J., 2012. The signaling power of sanctions in social dilemmas. Journal of Law, Economics, and Organization 28, 103–125.

von Siemens, F. A., 2013. Intention-based reciprocity and the hidden costs of control. Journal of Economic Behavior & Organization 92, 55–65.

Warneken, F., Tomasello, M., 2006. Altruistic helping in human infants and young chimpanzees. Science 311, 1301–1303.

Warneken, F., Tomasello, M., 2008. Extrinsic rewards undermine altruistic tendencies in 20-month-olds. Developmental psychology 44, 1785.

Weitzman, M.L., 1980. The "ratchet principle" and performance incentives. The Bell Journal of Economics 11, pp. 302–308.

Appendix A. Proof related to Section 2

Proof of Proposition 1. The producer's utility is $u(y) = p(y)y - c(y)$ with $u'(y) = p'(y)y + p(y) - c'(y)$ and $u''(y) = p''(y)y + 2p'(y) - c''(y)$. Given that marginal revenue is falling ($p''(y)y + 2p'(y) < 0$) and $c'' > 0$, the objective function is concave and the first-order condition is necessary and sufficient. By definition, $c'(y^\nu) = 0$ so that $u'(y^\nu) < 0$ if and only if $p'(y^\nu)y^\nu + p(y^\nu) < 0$. Re-arranging terms yields $1 - (-\frac{p'(y)y}{p(y)}) = 1 - |\epsilon| > 0$. Thus, $u'(y^\nu) < 0$ and maximal utility under tradeability is attained for some $y^\tau < y^\nu$ if and only if $\epsilon > -1$. □

Appendix B. Proofs related to Section 3

Lemma 2. *In any Perfect Bayesian Nash Equilibrium (PBNE), the consumer's second period price offer following any history in which the producer chose $y_1 = k$, $k \in \{0, 1\}$, denoted p_k, satisfies*

$$p_k \leq 1.$$

Proof. In any perfect Bayesian Nash equilibrium, the producer's second period choices must satisfy $y_2 = 0$ whenever $p_k < -\theta$ and $y_2 = 1$ whenever $p_k > -\theta$. If $p_k = -\theta$ he can make either choice. The consumer's expected utility from offering p_k is thus $(1 - G_k(-p_k)) \cdot (B - p_k)$, where G_k denotes the posterior distribution of θ given that $y_1 = k$. Given the prior distribution of types, $G_k(-p_k) = 0$ for all $p_k \geq 1$, and thus the consumer's expected utility is decreasing in p_k for $p_k \geq 1$. It follows that his best offer must be less than or equal to 1. □

Lemma 3. *In any perfect Bayesian Nash equilibrium, the producer's first period behavior follows a cutoff rule of the form $y_1 = 0$ if $\theta < \hat{\theta}$ and $y_1 = 1$ if $\theta > \hat{\theta}$, where $\hat{\theta} \in (-1, 1]$. (For $\theta = \hat{\theta}$, either action is possible.)*

Proof. Let p_k be the equilibrium period 2 price offer after a period 1 choice $y_1 = k$, $k \in \{0, 1\}$. By Lemma 2, we have $p_k \in [0, 1]$. Consider a producer of type $\theta \in [-1, +1]$. Choosing $y_1 = 1$ gives him θ today and $\max\{0, \theta + p_1\}$ tomorrow. Choosing $y_1 = 0$ gives him 0 today and $\max\{0, \theta + p_0\}$ tomorrow. The net benefit of choosing $y_1 = 1$ is

$$g(\theta) \equiv \theta + \max\{0, \theta + p_1\} - \max\{0, \theta + p_0\}.$$

A producer of type θ strictly prefers to choose $y_1 = 1$ if $g(\theta) > 0$, strictly prefers $y_1 = 0$ whenever $g(\theta) < 0$, and is indifferent when $g(\theta) = 0$. Using Lemma 2, we see that $g(-1) = -1$ and $g(1) = 1 + (p_1 - p_0) \geq 0$. By continuity, there exists $\hat{\theta} \in (-1, 1]$ such that $g(\hat{\theta}) = 0$. Next, we show by contradiction that there cannot be multiple indifferent producers. Suppose instead that $g(\tilde{\theta}) = g(\hat{\theta}) = 0$ for $\tilde{\theta} > \hat{\theta}$. Since $g(\theta)$ is weakly increasing, it follows that $g(\theta)$ is constant on the interval $(\hat{\theta}, \tilde{\theta})$. Then we must have $\theta + p_1 < 0 < \theta + p_0$ for all $\theta \in (\hat{\theta}, \tilde{\theta})$. Therefore $p_0 > p_1 \geq 0$. Further, $g(\theta) = \theta - (\theta + p_0) = -p_0 < 0$ for $\theta \in (\hat{\theta}, \tilde{\theta})$, which contradicts $g(\theta) = 0$. It follows that there exists a unique $\hat{\theta}$ such that $g(\hat{\theta}) = 0$. Further, $g(\theta) < 0$ for $\theta < \hat{\theta}$, and $g(\theta) > 0$ for $\theta > \hat{\theta}$. □

Proof of Lemma 1. We solve the model by backward induction beginning with the producer's second period activity choice. In period 2, the producer chooses $y_2 = 1$ if $\theta > -p$ and $y_2 = 0$ if $\theta < -p$. If $\theta = -p$ he may make either choice. Given this, the consumer's optimal offer p depends on her beliefs concerning the producer's type, θ. These beliefs, in turn, must be consistent with the producer's period 1 behavior and the cutoff rule established in Lemma 3. We proceed by considering separately the cases $y_1 = 0$ and $y_1 = 1$.

Case (i): $y_1 = 1$. In this case, $\theta \sim \mathrm{U}[\hat{\theta}, 1]$, with cumulative distribution function $G_1(\theta) = \frac{\theta - \hat{\theta}}{1 - \hat{\theta}}$. The producer will engage in the activity with probability $1 - G_1(-p) = \min\left\{1, \frac{1+p}{1-\hat{\theta}}\right\}$. As a function of p, the consumer's second period expected utility is given by

$$h_1(p) = \begin{cases} \left(\frac{1+p}{1-\hat{\theta}}\right) \cdot (B - p) \text{ if } p \leq -\hat{\theta} \\ 1 \cdot (B - p) \text{ if } p > -\hat{\theta} \end{cases}$$

If $\hat{\theta} \geq 0$, $h_1'(p) = -1$ for all $p \geq 0$ and so $p_1 = 0$ is the maximizing choice. Next consider $\hat{\theta} < 0$. Then $h_1'(p) = -1$ for $p \geq -\hat{\theta}$, so the maximum occurs at $p_1 \in [0, -\hat{\theta}]$. On that interval, $h_1'(p) = (B - 1 - 2p)/(1 - \hat{\theta})$. If $B < 1$, this is negative for all $p \geq 0$ and so $p_1 = 0$. If $B > 1 - 2\hat{\theta}$, the derivative is always positive and so $p_1 = -\hat{\theta}$. Otherwise, the maximum occurs inside the interval where $h_1'(p) = 0$, i.e., at $p_1 = (B - 1)/2$. To summarize, the consumer's optimal price offer after observing $y_1 = 1$ is given by

$$p_1 = \begin{cases} 0 \text{ if } \hat{\theta} \geq 0 \text{ or } B < 1 \\ \frac{B-1}{2} \text{ if } \hat{\theta} < 0 \text{ and } B \in [1, 1 - 2\hat{\theta}] \\ -\hat{\theta} \text{ if } \hat{\theta} < 0 \text{ and } B > 1 - 2\hat{\theta} \end{cases} \tag{B.1}$$

Case (ii): $y_1 = 0$. In this case, $\theta \sim \text{U}[-1, \hat{\theta}]$, with cumulative distribution function $G_0(\theta) = \frac{\theta+1}{\hat{\theta}+1}$. The producer will engage in the activity with probability $1 - G_0(-p) = \max\left\{0, \min\left\{1, \frac{\hat{\theta}+p}{\hat{\theta}+1}\right\}\right\}$. The consumer's second period expected utility is given by

$$h_0(p) = \begin{cases} 0 \text{ if } p < -\hat{\theta} \\ \left(\frac{\hat{\theta}+p}{\hat{\theta}+1}\right) \cdot (B - p) \text{ if } p \in [-\hat{\theta}, 1] \\ 1 \cdot (B - p) \text{ if } p > 1 \end{cases}$$

Observe that $h_0(p)$ is constant for $p < -\hat{\theta}$, strictly concave for $p \in [-\hat{\theta}, 1]$, and strictly decreasing for $p > 1$. For $p \in [-\hat{\theta}, 1]$, $h_0'(p) = (B - \hat{\theta} - 2p)/(\hat{\theta}+1)$. If $B > 2 + \hat{\theta}$, this is positive for all $p < 1$ and so the maximizing choice is $p_0 = 1$. If $B < -\hat{\theta}$, it is negative for all $p > -\hat{\theta}$ and so any $p_0 \leq \hat{\theta}$ is a maximizing choice. Otherwise, the maximum occurs inside the interval, where $h_0'(p) = 0$, i.e., at $p_0 = (B - \hat{\theta})/2$. To summarize, the consumer's optimal price offer after observing $y_1 = 0$ is given by

$$p_0 \in \begin{cases} \left[0, -\hat{\theta}\right] \text{ if } B < -\hat{\theta} \\ \left\{\frac{B-\hat{\theta}}{2}\right\} \text{ if } B \in [-\hat{\theta}, 2 + \hat{\theta}] \\ \{1\} \text{ if } B > 2 + \hat{\theta} \end{cases} \tag{B.2}$$

Next we use these optimal price offers to pin down the equilibrium cutoff strategy $\hat{\theta}$, by analyzing the producer's period 1 activity choice. Suppose first that $\hat{\theta} < 0$, i.e., that the indifferent producer type is not intrinsically motivated. In that case, his expected utility of engaging in the activity

in period 1 is $\hat{\theta}$ today, and 0 tomorrow. (The latter follows from the fact that $p_1 \leq -\hat{\theta}$.) Thus the expected utility of choosing $y_1 = 1$ is strictly negative. In contrast, the expected utility of choosing $y_1 = 0$ is zero today and $\max\{0, p_0 + \hat{\theta}\} \geq 0$ tomorrow and hence weakly positive. Thus an agent of type $\hat{\theta}$ strictly prefers not to engage in the activity in period 1, a contradiction.

It follows that $\hat{\theta} \geq 0$. Then $p_1 = 0$ and the indifferent producer's expected utility of choosing $y_1 = 1$ is $\hat{\theta}$ both today and tomorrow. Choosing $y_1 = 0$ gives zero today and $\hat{\theta} + p_0$ tomorrow, where $p_0 = \min\{1, \frac{B-\hat{\theta}}{2}\}$. The agent is indifferent if $2\hat{\theta} = \hat{\theta} + p_0$, i.e. if $\hat{\theta} = p_0 = \min\{1, \frac{B-\hat{\theta}}{2}\}$. The unique solution to this condition is given by

$$\hat{\theta} = \begin{cases} \frac{B}{3} \text{ if } B \in [0, 3) \\ 1 \text{ if } \quad 3 \leq B. \end{cases}$$

Given this, equations (B.1) and (B.2) can be used to verify that the consumer offers a price exactly equal to $\hat{\theta}$ if and only if the producer refrains from the activity in period 1. □