Smith, John

Working Paper

Cognitive dissonance, imperfect memory and the preference for increasing payments

Working Paper, No. 2007-05

Provided in Cooperation with:
Department of Economics, Rutgers University

Suggested Citation: Smith, John (2007) : Cognitive dissonance, imperfect memory and the preference for increasing payments, Working Paper, No. 2007-05, Rutgers University, Department of Economics, New Brunswick, NJ

This Version is available at:
https://hdl.handle.net/10419/31260

Cognitive Dissonance, Imperfect Memory and the Preference for Increasing Payments*

John Smith†
Rutgers University-Camden

August 13, 2007

Abstract

In this paper we propose a theory of cognitive dissonance through imperfect memory. Cognitive dissonance is the tendency of a person to engage in self justification after a decision. We offer an interpretation of the single decision cognitive dissonance experiments: an agent has an unknown cost of effort and before the decision receives a private signal of the cost of effort, which is subsequently forgotten. Following the decision, the agent makes an inference regarding the content of this signal based on the publicly available information: the action taken and the wage paid. We explore the implications of this interpretation in a setting requiring a decision of effort in two periods. A preference for increasing payments naturally emerges from our model. With the auxiliary assumption that obtaining wage income requires an unknown cost of effort and obtaining rental income requires a known, zero cost of effort, our results provide an explanation for the experimental findings of Loewenstein and Sicherman (1991). These authors find evidence of stronger preferences for increasing "income from wages" rather than "income from rent." Our model makes the novel prediction that this preference for increasing payments will only occur when the contracts are neither very likely nor very unlikely to cover the cost of effort.

*First Version: May 2005. This paper has benefited from discussions with Roland Benabou, Faruk Gul, Jo Hertel, Marcelo Pinheiro, participants of the Midwest Economics Association Meetings in Minneapolis, the SABE Conference in New York and the 2005 Whitebox Advisors Graduate Student Conference at Yale.

†Email: smithj@camden.rutgers.edu; Website: http://crab.rutgers.edu/~smithj; Phone: (856) 225-6319.

1 Introduction

A vast number of experiments have identified a tendency of subjects to engage in self-justification after making a decision. These experiments show that the extent of self-justification is negatively related to the material inducement for the decision. Specifically, those paid less for completion of a task later tend to report the task as more enjoyable than those paid more. This behavioral phenomenon is referred to as cognitive dissonance.

In this paper we propose a theoretical framework for modeling the behavior identified by the cognitive dissonance literature. An agent has an unknown cost of effort. Before the decision, the agent receives a private signal regarding this cost. Subsequently, the agent forgets the private signal but makes an inference regarding its content from the publicly available facts which we assume are recalled: the action taken and the wage. We show that behavior consistent with the cognitive dissonance experiments naturally emerges from these assumptions: a smaller amount of material surplus leaves the agent with a lower ex-ante estimation of the cost of effort.[1]

We then explore the implications of this decision problem where actions are chosen in two periods. We assume that in the second period the agent recalls that the signal has been forgotten and that in the first period the agent anticipates this outcome. It might then be advantageous for the agent in the first period to manipulate second period beliefs through first period actions. This manipulation is more easily achieved through payments which are increasing rather than constant. Thus our agent can simultaneously have a preference for increasing sequences of payments which require an unknown cost of effort and a preference for constant payments for which effort is known to be costless.

This implication of our model relates to the choice experiments found in Loewenstein and Sicherman (1991).[2] These authors find that people have a stronger preference for an increasing sequence of payments when described as "income for wages" rather than "income from rent." Ostensibly, effort is required for the acquisition of income from wages as opposed to income from rent.[3] We argue that this choice reflects an effort to reduce the perceived cost of effort. In other words, we contend that the findings of Loewenstein and Sicherman are the result of an optimal application of cognitive dissonance: the agent prefers to induce future cognitive dissonance, by getting paid less earlier and more later.

In order to illustrate the role of imperfect memory in our modeling of the baseline cognitive dissonance experiments, consider the following simple example.

Example 1 *Suppose an agent lives for two periods and is to take an action in the first period. The cost of effort for the action is imperfectly known to the agent. In the second period, the*

[1] For more on the relationship between this paper and the concepts of cognitive dissonance and Self-Perception Theory, see the discussion at the conclusion of the paper.

[2] Later confirmed by Gigliotti and Sopher (1997) and Matsumoto et. al. (2000).

[3] This interpretation seems to be consistent with the intent of the authors.

agent provides her updated beliefs about the cost of effort. Here we assume that the cost of effort can either be high or low: $c \in \{c_H, c_L\}$ *where* $c_L = 0$ *and* $c_H = 1$*. Before deciding on an action, the agent receives a signal (*$s \in \{s_H, s_L\}$*) which imperfectly reveals the true cost of effort:*

$$\begin{aligned} P(s_H|c_H) &= P(s_L|c_L) = q \\ P(s_L|c_H) &= P(s_H|c_L) = 1 - q \end{aligned}$$

where $q \in (0.5, 1)$*. The agent then decides to take the action or not (*$a \in \{1, 0\} = A$*) where* 1 *indicates high effort and* 0 *indicates low effort. The agent has the following ex-ante beliefs about the true cost of effort:* $\beta = P(c_H) = 0.5$ *and* $1 - \beta = P(c_L) = 0.5$*. In the second period, the agent can only recall the publicly available information. In other words the agent recalls* a *and* w *but not* s*. From the information which is recalled, the agent makes an inference regarding the true cost of effort. The posterior belief of a high cost of effort is denoted as* $\overline{\beta}(a, w(a))$*. A contract is a mapping from actions into payments (*$w : A \Rightarrow \mathbb{R}_+$*). Consider two different contracts* w *and* w'*. The contract* w *specifies that* $w(1) = z$ *and* $w(0) = 0$*. Whereas* w' *is such that* $w'(1) = z'$ *and* $w'(0) = 0$ *where* $z > z'$*. Further suppose that*

$$z > q > z' > 1 - q$$

Given this condition, under contract w *the agent will select* $a = 1$ *for any signal, however under contract* w' *the agent will select* $a = 1$ *if* $s = s_L$ *and* $a = 0$ *if* $s = s_L$*. It follows that* $\overline{\beta}(1, z) = 0.5 > \overline{\beta}(1, z') = 1 - q$ *despite the fact that* $z > z'$*. In other words, the contract paying more for completion of the task will produce worse beliefs about the task, conditional upon completion of the task.*■

This simple example serves to illustrate our interpretation of the baseline cognitive dissonance experiments. After completion of a task, the agent with smaller surplus will regard the task as less costly than an agent with a larger surplus. This is because the inference after the smaller payment rules out the receipt of the low signal whereas no such inference can be made after the larger contract.

In this paper we model an agent who acts as in Example 1 but makes a decision regarding effort in two periods. In the inference of past, forgotten signals, the agent engages in self-justification. Through this self-justification, the agent has a preference for increasing sequences payments. We provide a formal definition as such in a *Preference for Increasing Payments* (PIP). This definition provides a standard such that if we observe a choice of an increasing contract w^I over a constant contract w^C then we conclude that the agent has a genuine preference for increasing payments. By this we mean that if the conditions of PIP are satisfied then we can conclude that the choice could only have been made by a person with an intrinsic preference for increasing sequence of payments. Furthermore, if the conditions of PIP have been satisfied, then we have identified the behavior found by Loewenstein and Sicherman.

1.1 Related Literature: Imperfect Memory

There exists a literature which models agents with imperfect memory. This concept has been used in a variety of applications, however imperfect memory has not been used in connection with a preference for increasing payments. The reader interested in a careful discussion of the relationship between the imperfect memory literature and this paper should proceed to the balance of this subsection.

We are not the first to assign a different status of memory on the basis of the type of information. For instance, Mullainathan (2002) makes a distinction between "hard" information and "soft" information where it is assumed that hard information is perfectly recalled and soft information is not. Like the present paper, Mullainathan considers private information to be soft and publicly available information to be hard. Other examples of authors assuming that actions and payments are not forgotten and private signals are forgotten include Swank (2006) and Hirshleifer and Welch (2002).

We are not the first to model an agent with imperfect memory making inferences about personal characteristics through past actions. We are also not the first to assume that, subsequent to memory loss, the agent is aware that information as has been forgotten. In Benabou and Tirole (2004) a decision maker is both subject to imperfect recall and anticipates this imperfect recall. Benabou and Tirole assume that the agent is uncertain of the extent of their time inconsistency and makes an inference based on past actions. Benabou and Tirole assume that actions can be forgotten, however we assume that only private signals can be forgotten. Also by contrast Benabou and Tirole allow for the possibility that information will be forgotten with an interior probability, however we assume that the information is only forgotten with certainty. Indeed the mechanism of making inferences of past characteristics through past actions in order to model cognitive dissonance-like effects was suggested by Benabou and Tirole (2003).[4] The novelty of this aspect of our modeling technique lies in its application to a preference for increasing payments.

Similar to Bernheim and Thomadsen (2005), we model the agent with imperfect recall as relatively sophisticated and use a similar concept of optimality.[5] We assume that the agent can be modeled as several distinct players, each corresponding to the point in time in which their information is unique. In other words, there is an incarnation of the agent every time information has been gained or lost. This implies that, in any period, the agent can only deviate from the optimal strategy among the strategies available to him during that particular period.

Like our paper, Swank (2006) presents a two period model of imperfect recall. Primarily, what distinguishes Swank from our paper are the applications. Swank is concerned with

[4] Page 505, point d.

[5] Which the authors refer to as "modified multiself consistency" and attribute to Piccione and Rubenstein (1997).

examining the role of material incentives as supporting or discouraging intrinsic motivation. Whereas we are primarily concerned with the preference for increasing payments.

In Hirshleifer and Welch (2002) an agent tries to learn the true state of the world regarding the profitability of a project. The authors see their work as distinct from the cognitive dissonance literature[6] in part because they focus on the case where the agents maximize utility separately in each period. In other words, their agents do not exhibit inter-period strategic behavior. As such, their work is similar to the cascade literature. By contrast, in our model it is the inter-period strategic behavior which drives the results.

In this paper it is assumed that the imperfect memory comes in the form that the agent recalls the fact that the information has been forgotten. By contrast there is a literature which analyzes the issues which arise when such assumption is not made. Wilson (2004) models a decision problem of an agent with a limited memory capacity. Wilson assumes that agents cannot determine in which period each element of the memory capacity was obtained. Therefore, the agent does not know whether subsequent information has been forgotten. This assumption is less appropriate in our self-justification context. For self-justification, the agent must have *more* accurate recall of the action taken than of the private, personal the reasons for the action. For papers which concentrate on the implications of this form of imperfect recall, without explicitly modeling its mechanism, see the literature on both the Absent Minded Driver Problem[7] and the Sleeping Beauty Problem.[8]

Finally, Benabou and Tirole (2003) present a model where higher payments can reduce an agent's utility from undertaking a task. In their model, an informed principal offers a contract to an uninformed agent. Specifically, the principal knows the cost of effort and the agent does not. The agent makes an inference regarding the cost of effort from the contract offered. By contrast here we analyze a choice problem rather than a game between an informed principal and an uninformed agent. As in Loewenstein and Sicherman, we analyze the decision problem of an agent over various payment schemes.

1.2 Related Literature: Preference for Consistency

The notion of consistency is similar to the theory of cognitive dissonance. In fact, for some time it was thought that inconsistency produced the cognitive dissonance effects.[9] However, some psychologists do not view inconsistency as either a necessary or sufficient condition for cognitive dissonance.[10] What is agreed upon, however, is the structure of the experiments and their results. Every cognitive dissonance experiment contains an action and a subsequent reflection on that action. Further, this reflection is inconsistent with standard reward based theories, it is often associated with psychological arousal and it is in this sense that cognitive

[6] Page 403.
[7] Piccione and Rubenstein (1997) along with the balance of the issue of Games and Economic Behavior.
[8] Elga (2000), Lewis (2001), Monton (2002), Dorr (2002), Bradley (2003) and Weintraub (2004).
[9] For more on this see Smith (2007) and a general account of the topic see Harmon-Jones and Mills (1999).
[10] See Cooper (1999).

dissonance experiments can be described as self-justification. For instance, a typical cognitive dissonance experiment would pay an subject for completion of a task. After completion of the task, the subjects paid less would typically regard the task as more enjoyable than the subjects paid more.

In this paper, we model cognitive dissonance as self-justification through imperfect memory and we show that these assumptions imply a preference for increasing payments. The reader interested in a careful discussion of the relationship between the consistency literature and this paper should proceed to the balance of this subsection.[11,12]

Yariv (2006) models an agent with a preference for consistent beliefs by incorporating this as a separate term in the utility function. Therefore, the agent will make a trade-off between material payoffs and consistency payoffs in selecting actions and beliefs. We distinguish between Yariv and this paper by noting that we do not assume a preference for consistency. Here the agent acts as if seeking to achieve consistency, but strives for the optimal action using the publicly available information in interpreting the past action. And so in this paper, the beliefs of the agent are rational in that they are entirely determined by the objective features of the situation and are not a matter of choice.

Eyster (2002) models an agent with a preference to avoid regret. This preference is incorporated into the utility function of the agent. The effort to avoid regret induces behavior as if the agent has a preference for consistency. A difference between Eyster and this paper is that Eyster requires actions to be ex-post suboptimal in order for the identified effect to occur. Specifically uncertainty must be resolved in the determination of the regret. However, in this paper uncertainty is not resolved prior to the decision in either period. This should not be surprising as we are interested interpreting experiments in which a choice is made prior to any resolution of uncertainty.

Finally, we interpret the cognitive dissonance literature differently than do Epstein and Kopylov (2006). These authors assume that the agent perfectly anticipates cognitive dissonance and prefers to avoid this effect. Specifically, they assume that the agent is tempted by the cognitive dissonance inducing choice and prefers a menu without such an option. By contrast, in our model the optimal choice of the agent will often induce cognitive dissonance.

2 Model

The previous section demonstrated our interpretation of the cognitive dissonance experiments involving a single decision. Now we explore the implications of imperfect memory in a setting involving two decisions.

[11] A discussion of the relationship between the model presented here and Self-Perception Theory is offered at the end of the paper.

[12] For models which assume cognitive dissonance and examine its implications, see Akerlof and Dickens (1982), Konow (2000) and Oxoby (2003, 2004).

We assume an agent with standard and separable preferences with regard to money. Utility for money ($u : \mathbb{R}_+ \Rightarrow \mathbb{R}_+$) is everywhere increasing, concave and differentiable.

An agent is to complete a task in two periods. In periods 1 and 2 the agent decides to take the action or not ($\{1, 0\} = A$). A contract[13] is a mapping from actions into payments ($w_1 : A_1 \Rightarrow \mathbb{R}_+$ and $w_2 : A_2 \Rightarrow \mathbb{R}_+$). For simplicity we assume that $w_t(0) = 0$. Therefore, we can summarize the contract w by the pair of payments rendered for high effort (w_1, w_2). Lower case w will denote single contract and upper case W will denote a menu of contracts.

The cost of low effort is known to be 0. The cost of high effort is unknown. Both the agent and the principal have identical ex-ante beliefs about the true cost of effort. Priors regarding the cost of effort c are equally distributed on $\{c_L, c_H\} = C$. Further we assume that $c_H = 1$ and $c_L = 0$.

Before deciding the effort level in periods 1 and 2, the agent receives a signal $s \in C$, which imperfectly reveals the true cost of effort. We assume that the signal is as follows:

$$P(s_j|c_i) = \begin{cases} q \text{ if } c_i = s_j \\ (1-q) \text{ if } c_i \neq s_j \end{cases}$$

where $q \in (0.5, 1)$. Conditional on c, we assume that the signals s_1 and s_2 are statistically independent.

As stated above, before the decision is made, the agent receives an imperfect signal of the cost of effort. However, in the subsequent period, the agent does not recall this signal and makes an inference regarding its content. In period 2 the agent will make an inference regarding s_1 from all available information: the value of a_1 (either 0 or 1) in response to w_1 and w_2 and the second period signal s_2. The second period agent cannot condition on s_1 as it has been forgotten.

The first and second period actions (a_1 and a_2) are mappings from all available information into a probability distribution on $\{0, 1\}$. The first period strategy can be written as $a_1(s_1, w_1, w_2)$ and the second period strategy can be written as $a_2(s_2, w_2, a_1(w_1, w_2))$. Our notation reflects the above discussion as we show the dependence of a_2 on the wages w_1, w_2, the action selected in the first period a_1 and second period signal s_2, but not the first period signal s_1. Where there is no risk of confusion, we suppress the redundant notation in $a_1(\cdot)$ and $a_2(\cdot)$. Also note that $a_1(c_H)$ and $a_1(c_L)$ (both $\in [0, 1]$) refer to first period strategies and a_1 ($\in \{0, 1\}$) refers to the action actually taken.

To summarize the timing: in period 0 the agent selects a contract among those available. In periods 1 and 2 the agent determines whether to exert high or low effort given all information known during that period.

[13] Although we refer to these objects as 'contracts' this should not be interpreted as suggesting that we are analyzing strategic issues in a principal-agent setting. Here we exclusively focus on the choice problem of the agent in the experiments of Loewenstein and Sicherman (1991).

Now we provide the expected utility functions of the agent in each period . The expected utility of the agent in the second period, after receiving s_2, observing a_1 and selecting a_2, can be written as:[14]

$$U_2(s_2, a_1) = a_2(s_2, a_1)\{u(w_2) - E_2[c|s_2, a_1]\} \tag{1}$$

Here in expression (1), the second period agent can only observe the second period signal s_2 and the value of a_1 as either 0 or 1, therefore we write $E_2[c|s_2, a_1]$.

The expected utility of the agent in the first period, after receiving s_1 and selecting a_1, can be written as:

$$\begin{aligned} U_1(s_1) &= a_1(s_1)\{u(w_1) - E_1[c|s_1]\} \\ &+ \sum_{s_2 \in C} P(s_2|s_1)[a_1(s_1)a_2(s_2, a_1 = 1)\{u(w_2) - E_2[c|s_2, a_1 = 1]\} \\ &+ (1 - a_1(s_1))a_2(s_2, a_1 = 0)\{u(w_2) - E_2[c|s_2, a_1 = 0]\}] \end{aligned} \tag{2}$$

Here in expression (2), the agent knows s_1 and knows the function $E_2[c|s_2, a_1]$ for each possible s_2. Furthermore, the agent in the first period uses s_1 to improve the prediction of s_2 through $P(s_2|s_1)$. It is worth noting that although the agent in the first period knows s_1, it is also known that the information will be forgotten in period 2. Therefore this sophistication assumption requires that the first period consideration of second period utility should include $E_2[c|s_2, a_1]$ rather than $E_2[c|s_2, s_1]$. If this was not the case, the agent would be naive about the upcoming imperfect recall.

We write the expected utility of the contract w in the ex-ante period as:

$$\begin{aligned} U_0(w) &= \frac{1}{2} \sum_{s_1 \in C} [a_1(s_1)\{u(w_1) - E_1[c|s_1]\} \\ &+ \sum_{s_2 \in C} P(s_2|s_1)[a_1(s_1)a_2(s_2, a_1 = 1)\{u(w_2) - E_2[c|s_2, a_1 = 1]\} \\ &+ (1 - a_1(s_1))a_2(s_2, a_1 = 0)\{u(w_2) - E_2[c|s_2, a_1 = 0]\}] \end{aligned} \tag{3}$$

Here in expression (3), the utility of the agent is the difference between the expected utility of money in each period and the expected cost of effort in that period.

For convenience in the upcoming definitions, we denote the probability that the agent selects high effort in period t as Ψ_t. Therefore we can write

$$\Psi_1 = \frac{a_1^*(c_L) + a_2^*(c_H)}{2}$$

[14] As these posterior beliefs are somewhat nonstandard, see the appendix for a more complete description.

and

$$\Psi_2 = \frac{1}{2}\{a_2^*(c_L, a_1 = 1)(qa_1^*(c_L) + (1-q)a_1^*(c_H))$$
$$+a_2^*(c_L, a_1 = 0)(q(1-a_1^*(c_L)) + (1-q)(1-a_1^*(c_H)))$$
$$+a_2^*(c_H, a_1 = 1)((1-q)a_1^*(c_L) + qa_1^*(c_H))$$
$$+a_2^*(c_H, a_1 = 0)((1-q)(1-a_1^*(c_L)) + q(1-a_1^*(c_H)))$$

We make two sophistication assumptions regarding the memory of the agent. We assume that the second period agent is sophisticated in remembering that the information has been forgotten. We also assume that the ex-ante and first period agents are sophisticated in that they anticipate this outcome. Although we regard these as strong assumptions, we do think that some sophistication is reasonable. While sophistication is necessary for the results in this paper, we do not expect the results to qualitatively change if the agent is only partially sophisticated.

We now list the conditions for optimal behavior. The following conditions require that at each period the agent maximize expected utility given what is known at the time. Specifically, we require that (Condition (i)) the the ex-ante player selects the contract among the menu of contracts which will yield the highest expected utility. We require that (Condition (ii)) the first period agent maximizes expected utility given the signal s_1. We require that (Condition (iii)) the second period agent maximizes expected utility given signal s_2 and the behavior of the first period agent a_1. We also assume that (without loss of generality) when indifferent between selecting high and low effort, the agent selects low.

Conditions (i) $w \in W$ such that $U_0(w) \geq U_0(w')$ for any $w' \in W$

(ii) $a_1^*(c_H)$ such that $U_1(c_H, a_1^*(c_H)) \geq U_1(c_H, a_1(c_H))$ for any $a_1(c_H)$ and $a_1^*(c_L)$ such that $U_1(c_L, a_1^*(c_L)) \geq U_1(c_L, a_1(c_L))$ for any $a_1^*(c_L)$.

(iii)

$$a_2^*(s_2, a_1) = \begin{cases} 1 \text{ if } u(w_2) > E_2[c|s_2, a_1] \\ 0 \text{ if } u(w_2) \leq E_2[c|s_2, a_1] \end{cases}$$

Conditions (i), (ii) and (iii) constitute our optimality requirements for the agent in the ex-ante period, period 1 and period 2 respectively. Note that Condition (iii) is without loss of generality as any tie breaking rule would not qualitatively change the following results. Also note that in Condition (ii) we interpret the first period agent receiving $s_1 = c_H$ and the first period agent receiving $s_1 = c_L$ as distinct players in a noncooperative game. The player only considers the payoffs conditioning on the signal actually received. In other words, under condition (ii) the first period player maximizes utility conditioning on the signal received even though this signal is to be forgotten with certainty: the agent does not consider how his strategy affects payoffs had the other signal been realized.

We now provide some further restrictions on the behavior we wish to consider. The following condition requires that the strategies are monotonic in the signal of the cost of effort.

Condition (M) $a_1^*(c_L) \geq a_1^*(c_H)$

Condition (M) helps to eliminate from consideration, signaling outcomes which we consider to be inappropriate in the context of understanding cognitive dissonance. For instance, a violation of Condition (M) would imply that $E[c|a_1 = 0] < E[c|a_1 = 1]$. We wish to avoid such counterintuitive signalling outcomes.

The next condition specifies the out-of-equilibrium beliefs after events with zero probability. It is assumed that if the first period strategy specifies that the agent always selects 1, however 0 is observed, then the agent infers that the agent received c_H. A similar condition applies to the case where the strategy specifies that 0 is always selected. We now specify this formally.

Condition (O) Out-of-equilibrium beliefs are such that if $a_1^*(c_H) = a_1^*(c_L) = 1$ and $a_1 = 0$ then period 2 agent infers that $s_1 = c_H$ with probability one. If $a_1^*(c_H) = a_1^*(c_L) = 0$ and $a_1 = 1$ then period 2 agent infers that $s_1 = c_L$ with probability one.

Conditions (M) and (O) might strike the reader as rather blunt instruments. On the other hand, we are considering a setting with non standard posterior beliefs and the possibility of unusual signalling outcomes. Therefore, we seek to make our assumptions as simple and transparent as possible. Alternatively, we could weaken these assumptions, however we would be obliged to consider optimal behavior which we do not consider to be helpful in understanding self-justification and the preference for increasing payments.

We now define our notion of optimality:

Definition 1 *The agent is* ***Self-Justification Optimal*** *(SJO) if Conditions* (i), (ii), (iii), M *and* O *are satisfied.*

We conclude this section with the following result which illustrates an important implication of optimality: an agent cannot mix at the same rate after c_H and c_L, under either definition of optimality. This result is significant as it captures the key insight that the first period agent exchanges current payoffs for an improvement in future beliefs. If there is no future benefit, as the inference after each action is identical, then this exchange is not undertaken The following proposition formalizes this statement.

Proposition 1 *It cannot be an SJO that* $a_1^*(c_H) = a_1^*(c_L) \in (0, 1)$

Proof: See appendix.

The intuition behind the result is as follows: the decision at period 1 is influenced by the anticipated posteriors in period 2. Particularly, if the first period agent selects an action

other than the myopically optimal one then there must be some future benefit in the form of improved posteriors. If mixing occurs after both signals then it is certain that a myopically suboptimal action is taken. However, if the agent mixes at the same rate after both c_L and c_H then the posteriors are unaffected and so this action will not produce a benefit. Therefore, identical mixing after c_L and c_H is inconsistent with optimality.

3 A Preference for Increasing Payments

In this section we discuss the main implication of the model: an agent with imperfect memory can display a preference for an increasing sequence of payments. To make this notion more precise we offer the definition of a *Preference for Increasing Payments (PIP)*.

Consider a constant contract w^C. PIP places restrictions on a candidate increasing contract w^I such that if w^I is preferred over w^C then we are able to conclude that this choice is the result of an intrinsic preference for increasing payments. This intrinsic preference is the result of the self-justification through imperfect memory. Operationally, PIP requires that the candidate increasing contract w^I is less *valuable* than the constant contract w^C. Additionally w^I must be *close* to w^C in that both contracts induce qualitatively similar behavior. If w^I satisfies these requirements of valuableness and closeness with respect to a w^C and additionally w^I is preferred over w^C then we conclude that the agent has an intrinsic preference for increasing payments.

The utility of PIP comes in isolating the type of behavior found in Loewenstein and Sicherman (1991). Recall that Loewenstein and Sicherman find that people have a stronger preference for increasing payments of "income from wages" rather than "income from rent." We interpret "income from wages" as requiring an unknown cost of effort and "income from rent" as requiring a cost of effort known to be zero. Such a finding supports our model. Further, such findings cannot be explained by the class of contracts which we eliminate from consideration through the following definition. Rather, the following definition serves to isolate the role of self-justification in the preference for increasing payments. By placing these restrictions on the contracts, we are assured that we will declare that an agent exhibits a preference for increasing payments only when we observe behavior consistent with Loewenstein and Sicherman.

Before we provide our definition of the preference for increasing payments, we state our goals for the definition. Suppose that an agent is deciding between a constant contract w^C and an increasing contract w^I. If the benefits of the revenue from w^I exceeds the benefits of the revenue from w^C, and the agent prefers w^I, we would be wrong to conclude that the agent exhibits a preference for increasing payments. The agent simply prefers more money to less. Therefore, in determining whether an agent exhibits a preference for increasing payments, we will require that the expected revenue from w^C exceed that from w^I. We refer to this requirement as w^I being less *valuable* than w^C.

Definition 2 *A contract w^I is less **valuable** than contract w^C if*

$$\begin{aligned} & u(w_1^C)\Psi_1^C + u(w_2^C)\Psi_2^C \\ > \quad & u(w_1^I)\Psi_1^I + u(w_2^I)\Psi_2^I \end{aligned} \tag{4}$$

Now suppose that w^I is less valuable than w^C. However, suppose that w^C and w^I are sufficiently dissimilar so that qualitatively different behavior is induced. Again, we would be wrong to conclude that the agent exhibited a preference for increasing payments. We would only be justified in concluding that the agent likes the overall characteristics of w^I more than those of w^C. In particular we want the optimal second period strategies for w^I and w^C to be identical. We also want $u(w_1^I)$ and $u(w_1^C)$ to both be members of the same open interval, where such membership implies qualitatively similar behavior. We refer to this requirement as w^I being *close* to w^C.

Explicitly we define the following intervals as:

$$\begin{aligned} \Gamma_1 &: = (1-q, \frac{1}{2}) \\ \Gamma_2 &: = (\frac{1}{2}, \lambda^*) \\ \Gamma_3 &: = (\lambda^*, q) \\ \Gamma_4 &: = (q, \mu^*) \\ \Gamma_5 &: = (\mu^*, \frac{q^2}{q^2+(1-q)^2}) \\ \Gamma_6 &: = (\frac{q^2}{q^2+(1-q)^2}, \infty) \end{aligned}$$

where $\cup_{i=1}^6 \Gamma_i = \Gamma$ and $\frac{1}{2} < \lambda^* < q < \mu^* < \frac{q^2}{q^2+(1-q)^2}$. Note that the specific values of λ^* and μ^* are determined by the contract w^C and the value of q.[15] These intervals are selected in order to ensure that when comparing contracts, based on the idiosyncrasies of the information structure, we do not expect qualitatively different behavior to be induced by w^C and w^I. In particular, closeness requires that the first period payments of both contracts fall into a single Γ_j as listed above (Expression (5)) and that second period behavior is identical for both contracts (Expression (6)).

Definition 3 *Contracts w^C and w^I are **close** if*

$$u(w_1^C) \in \Gamma_i \text{ if and only if } u(w_1^I) \in \Gamma_i \text{ for some } \Gamma_i \in \Gamma \tag{5}$$

and

$$a_2^{*C}(s_2, a_1) = a_2^{*I}(s_2, a_1) \text{ for } s_2 \in \{c_L, c_H\} \text{ and } a_1 \in \{0, 1\} \tag{6}$$

[15] For more on λ^* and μ^* see the proofs of Lemmas 5 and 6 respectively in the appendix.

We now state the definition which provides the criteria for determining whether the agent displays a preference for increasing payments of the type found in the self-justification choice experiments. We require that for such a determination there must exist an increasing contract which is less *valuable* than an increasing contract and that the two contracts are *close*. If these two conditions hold and the increasing contract is preferred over the constant contract then we declare that the agent exhibits a preference for increasing payments.

Definition 4 *Consider a constant contract w^C. An SJO agent displays a* ***preference for increasing payments*** *(**PIP**) if there exists an increasing contract w^I which is less valuable than w^C, close to w^C and*

$$U_0^*(w^I) > U_0^*(w^C)$$

Definition 4 provides a standard for determining whether the agent displays behavior consistent with our formulation of cognitive dissonance involving two decisions: the agent has a preference for increasing payments. Intuitively, if contract w^I pays more than contract w^C and w^C is similar to w^I and yet the agent prefers w^I to w^C then we say that the agent has an intrinsic preference for increasing payments. The conditions valuable and close serve to isolate the effects of the payments of the contracts for the agent with perfect memory.

To better understand the content of Definition 4 we provide the following example.

Example 2 *Assume that the agent perfectly recalls the signal and that $q = 0.7$. Contract w^C pays $u(w_1^C) = u(w_2^C) = 0.5$. Contract w^I pays $u(w_1^I) = 0.49$ and $u(w_2^I) = 0.51$. Under w^C the agent selects $a_1 = 1$ only after $s_1 = c_L$ and in the second period $a_2 = 1$ only after $s_2 = c_L$ and $s_1 = c_L$. Therefore total utility from contract w^C is then*

$$U_0(w^C) = 0.5[0.5 - 0.3 + 0.7[0.5 - \frac{0.09}{0.58}]] = 0.221$$

Under contract w^I the agent again selects high effort only after $s_1 = c_L$ however in the second period selects high effort for every signal pair other than $s_2 = c_H$ and $s_1 = c_H$. Therefore the total utility from contract w^I is then:

$$U_0(w^I) = 0.5[0.49 - 0.3 + 0.7[0.51 - \frac{0.09}{0.58}] + 0.3[0.51 - 0.5]] + 0.5[0.3[0.51 - 0.5]] = 0.222$$

■

A few aspects of Example 2 are worth noting. The first is that the second period posteriors are not close in that the two contracts induce different second period behavior. In Example 2, the increasing contract exploits the particulars of the information structure in order to become more attractive than the constant contract. This exploitation is indicated by the different second period behavior induced by the two contracts. Therefore, in determining whether an agent has a genuine preference for increasing payments, we will require that when comparing two contracts, the agent has identical second period behavior. The second notable aspect of the example relates to the assumption of perfect recall. In what follows, we show that an agent with perfect recall cannot exhibit PIP (Proposition 5).

4 SJO Behavior

We now provide an explicit characterization of the relationship between the parameter values and the exhibition of PIP when SJO is the optimality requirement. For a more complete characterization of SJO see Lemmas 4 through 7 in the appendix.

Together the propositions below produce a novel implication of our model: an agent displays PIP when it is not the case that the payments are very likely to cover the cost of high effort and it is not the case that the payments are very unlikely to cover the cost of high effort. In these cases the agent will find it worthwhile to seek a reduction in the perceived cost of effort by accepting a close and less valuable contract which induces more favorable beliefs.

In determining whether an agent displays PIP with respect to a constant contract w^C one needs to find an increasing contract w^I such that $w_1^I < w^C < w_2^I$. The open sets which appear in the following propositions allow us the possibility of locating such an appropriate w^I.

We start out with two negative results. The content of these two propositions can be summarized by the following: anytime behavior is identical between two close contracts the more valuable one will be preferred. Indeed, this is the content of Proposition 6 in the following section. Contracts in the regions described below will always induce constant behavior. We provide the parameter values in order to facilitate our discussion about the relationship between the nature of the uncertainty of the cost of effort and the resulting behavior.

Proposition 2 *If* (i) $u(w_1^C) = u(w_2^C) \in (1-q, \frac{1}{2})$ *or* (ii) $u(w_1^C) = u(w_2^C) > \frac{q^2}{q^2+(1-q)^2}$ *then an SJO agent never displays PIP.*

Proof: See appendix.

Proposition 3 (i) *If* $u(w_1^C) = u(w_2^C) \in (\frac{1}{2}, q)$ *then there is a* $\lambda^* \in (\frac{1}{2}, q)$ *such that if* $u(w_1^C) = u(w_2^C) \in (\frac{1}{2}, \lambda^*)$ *then the SJO agent never displays PIP.*

(ii) *If* $u(w_1^C) = u(w_2^C) \in (q, \frac{q^2}{q^2+(1-q)^2})$ *then there is a* $\mu^* \in (q, \frac{q^2}{q^2+(1-q)^2})$ *such that if* $u(w_1^C) = u(w_2^C) \in (\mu^*, \frac{q^2}{q^2+(1-q)^2})$ *then the SJO agent never displays PIP.*

Proof: See appendix.

The intuition is that in these regions, behavior is constant and therefore (see Proposition 6 in the following section) these contracts cannot constitute a PIP. Note the parameter values in the above propositions. There are two possibilities: either it is very unlikely that the contract will cover a high cost of effort (Propositions 2 (i) and 3 (i)) or it is very likely that the contract will cover a high cost of effort (Propositions 2 (ii) and 3 (ii)). In these cases the agent cannot satisfy PIP. It seems natural that in these cases the agent will not

find it worthwhile to seek a reduction in the overall cost of effort by selecting a less valuable contract.

Although the above two results are negative we now provide a positive result. The following proposition states that any time an agent is considering a contract w^C where $a_1^{*C}(c_H) \in (0,1)$ and $a_1^{*C}(c_L) = 1$ then it is always possible to find a less valuable and close increasing contract which is preferred over the constant contract. In other words, Proposition 4 says that the agent always displays PIP when $a_1^{*C}(c_H) \in (0,1)$ and $a_1^{*C}(c_L) = 1$.

Proposition 4 *(i) If* $u(w_1^C) = u(w_2^C) \in (\frac{1}{2}, q)$ *then there is a* $\lambda^* \in (\frac{1}{2}, q)$ *such that if* $u(w_1^C) = u(w_2^C) \in (\lambda^*, q)$ *then the SJO agent always displays PIP.*

(ii) If $u(w_1^C) = u(w_2^C) \in (q, \frac{q^2}{q^2+(1-q)^2})$ *then there is a* $\mu^* \in [q, \frac{q^2}{q^2+(1-q)^2})$ *such that if* $q < \mu^*$ *and* $u(w_1^C) = u(w_2^C) \in (q, \mu^*)$ *then the SJO agent always displays PIP.*

Proof: See appendix.

Although the proof of Proposition 4 is involved, the intuition is straightforward. Anytime $a_1^{*C}(c_H) \in (0,1)$ and $a_1^{*C}(c_L) = 1$, the ex-ante player prefers the first period player to select a smaller $a_1(c_H)$. It is always possible to find a close and less valuable increasing contract which induces a sufficiently smaller $a_1^*(c_H)$ in order to more than compensate for the smaller value of the increasing contract. Therefore, the agent satisfies PIP. Note the different wording for parts (i) and (ii) of Proposition 4. This is required as there exist parameter values such that $u(w_1^C) = u(w_2^C)$ and $q = \mu^*$.

The proposition provides an explicit characterization of the parameter values which induce this condition. Roughly, Proposition 4 says that if payment for the task neither very likely nor very unlikely to cover a high cost of effort, then the agent will exhibit PIP. This is a novel implication of our model.

The following proposition demonstrates the necessity of the imperfect recall assumption in our cognitive dissonance modeling.

Proposition 5 *If an agent has perfect recall of the signal of the cost of effort then the agent cannot display PIP.*

Proof: See appendix.

The intuition of the result is as follows. With perfect memory, first period actions do not affect second period beliefs. As a result, the first period action is not selected in order to influence second period posteriors. When the valuable condition is met then it must be that $U_0(w^C) - U_0(w^I) < 0$. Therefore the agent with perfect memory does not display PIP. Although the result appears to be straightforward, we hope to convince the reader of its significance as Proposition 5 supports our proposal of modeling the self-justification choice experiments through imperfect memory.

To give the reader a feel for the different mechanics in SJO we provide the following example.

Example 3 *Suppose that $q = 0.7$. The agent is considering two contracts. Contract w^C pays $u(w^C) = 0.6$. Contract w^I pays $u(w_1^I) = 0.59$ and $u(w_2^I) = 0.609$.*[16] *It follows that for both contracts*

$$a_2^*(c_L, a_1 = 1) = a_2^*(c_L, a_1 = 0) = 1$$
$$a_2^*(c_H, a_1 = 0) = 0$$

It is SJO under both contracts that $a_1^*(c_L) = 1$ for every choice of $a_1(c_H)$. The first period player receiving $s_1 = c_H$ then will seek to maximize:

$$U_1(c_H) = a_1(c_H)\{u(w_1) - 0.7\}$$
$$+0.7a_1(c_H)a_2(s_H, a_1 = 1)\{u(w_2) - E[c|c_H, a_1 = 1]\}$$
$$+0.3a_1(c_H)\{u(w_2) - E[c|c_L, a_1 = 1]\}$$
$$+0.3(1 - a_1(c_H))\{u(w_2) - E[c|c_L, a_1 = 0]\}$$

Contract w^C induces $a_1^{*C}(c_H) = 0.109$ and $a_2^{*C}(s_H, a_1 = 1) = 1$. This implies that $U_1^*(w^C, c_H) = 0.0338$, $U_1^*(w^C, c_L) = 0.610$ and an ex-ante utility of $2U_0^*(w^C) = 0.644$. Contract w^I induces $a_1^{*I}(c_H) = 0.102$ and $a_2^{*I}(s_H, a_1 = 1) = 1$. This implies that $U_1^*(w^I, c_H) = 0.036$, $U_1^*(w^I, c_L) = 0.611$ and an ex-ante utility of $2U_0^*(w^I) = 0.647$. Therefore the ex-ante player prefers w^I to w^C. To see valuable requirement is satisfied:

$$\begin{aligned} & 0.6(0.5)(1.0109) + 0.6[1 - (0.5)(0.7)(1 - 0.109)] \\ > \; & 0.59(0.5)(1.102) + 0.607[1 - (0.5)(0.7)(1 - 0.102)] \end{aligned} \tag{7}$$

Note also that close requirement is satisfied. Therefore the SJO agent exhibits PIP.■

The above example provides an opportunity to demonstrate some additional intuition behind the definitions of 'valuable' and 'close.' In Example 3, the contracts are such that $u(w_1^I) = 0.59$ $u(w_2^I) = 0.609$ and $u(w_1^C) = u(w_2^C) = 0.6$. Although it seems natural to conclude that the pairs are similar, one can confirm that the requirements for close are satisfied: both induce identical second period actions and both fall within the same open interval $\Gamma_i \in \Gamma$. Expression (7) demonstrates the valuable requirement. Therefore, an agent with perfect memory prefers the constant contract to the increasing contract.

We now provide a proposition which states that the exhibition of PIP is only generated by behavior.

Proposition 6 *Consider contracts w^I and w^C. If first period behavior is identical then PIP cannot occur.*

[16] Suppose that $u(w) = w^{\frac{1}{1.5}}$. Therefore $w_1^C = w_2^C = 0.465$, $w_1^I = 0.453$, $w_2^I = 0.475$ and $w_1^C + w_2^C > w_1^I + w_2^I$. In other words here again the agent actually prefers a smaller total payment so long as it is increasing.

Proof: See appendix.

The intuition behind Proposition 6 is as follows: the exhibition of PIP requires different posteriors for the two contracts and affecting the posteriors can only be achieved through the application different first period strategies for each contract. If the contracts do not induce such behavior then an increasing contract will never fulfill the requirements in PIP. Apart from providing intuition, Proposition 6 also has practical relevance. Suppose that we have identified the optimal strategies for two contracts. If the strategies are identical then the contract cannot be part of an exhibition of PIP.

5 Discussion

5.1 Conclusion

In this paper we have modeled cognitive dissonance through self-justification and imperfect memory. In particular, we extend the insights from the single decision psychology experiments into a setting requiring a decision in two periods. Our model aids in the interpretation of existing choice experiments which suggest that people have a stronger preference for increasing sequences of payments when described as "income from wages" rather than "income from rent." With the auxiliary assumption that there is an unknown cost of effort in obtaining the former and not the latter, our model provides a mechanism through which this behavior occurs. Our model not only replicates the Loewenstein and Sicherman results but also makes the additional prediction that such behavior will only be associated with contracts which are neither very likely nor very unlikely to cover a high cost of effort.

There remain several issues which will be the focus of future work. For instance, we are not clear about the significance of our choice of state space. A particularly interesting venture would seek to learn the behavior of our agent with a continuous state space. We are also eager to learn the weight of our sophistication requirements. Although we will need some sophistication for the results to hold, future work will examine the relationship between these weaker assumptions and behavior.

5.2 Cognitive Dissonance and Self-Perception Theory

It is possible that a reader will argue that the paper does not model cognitive dissonance, but rather the related concept of Self-Perception Theory as proposed by Bem (1972). Self-Perception Theory contends that the effects found in the cognitive dissonance experiments are due to a "cold" attempt to infer past attitudes from past actions rather than the psychological arousal associated with cognitive dissonance. Many psychologists consider Self-Perception Theory to be an incomplete subset of cognitive dissonance. In justification of this position, Zanna and Cooper (1974) perform an experiment where the subjects are allowed to misattribute the effects of the dissonance, by telling some subjects that a pill (actually a placebo) is responsible for any feelings of anxiety or psychological arousal. The authors find that those

given the means of misattribution did not adjust their beliefs in a manner consistent with the classic cognitive dissonance experiments. The authors conclude that self-perception theory cannot explain these results and therefore Self-Perception Theory cannot account for all of the effects in the cognitive dissonance experiments.

By contrast, our model can accommodate the misattribution described above. If the agent receives a signal indicating a costly level of effort, however believes that this signal has resulted from conditions other than a high true cost of effort (for instance, the effects of the pill) our agent will act as described in Zanna and Cooper (1974). For expositional clarity, we assumed a close relationship between the signals and the cost of effort. However, if such a relationship was not assumed, then the misattribution observed by Zanna and Cooper could be observed. In other words, the misattribution paradigm poses no particular problem to our model. Therefore, we maintain that our model provides an accurate description of the cognitive dissonance phenomena.

6 Appendix

The appendix is arranged as follows. In the first subsection, we first derive the posterior beliefs of the agent. These require attention as they are somewhat nonstandard. Then we provide a more complete description of the utilities of the agent. In the following subsection, we prove the lemmas which will be useful later. In the final subsection, we prove the results found in the body of the paper. We present the proofs of the results in the order to best facilitate their elucidation rather than the order presented in the body of the paper.

6.1 Preliminaries

We now derive the posterior beliefs of the period 2 agent given s_2 and $a_1 = 1$ as found in expression (1). If it is not the case that either $a_1(c_L) = a_1(c_H) = 1$ or $a_1(c_L) = a_1(c_H) = 0$ then

$$P(\widehat{c}|s_2, a_1) = \frac{P(s_2|\widehat{c})P(a_1|\widehat{c})}{\sum_{c\in C} P(s_2|c)P(a_1|c)}$$

since s_1 and s_2 are independent conditional on c. More explicitly

$$P(c|s_2 = c_H, a_1 = 1) = \begin{cases} \frac{q(qa_1(c_H)+(1-q)a_1(c_L))}{(q^2+(1-q)^2)a_1(c_H)+2q(1-q)a_1(c_L)} & \text{if } c = c_H \\ \frac{(1-q)((1-q)a_1(c_H)+qa_1(c_L))}{(q^2+(1-q)^2)a_1(c_H)+2q(1-q)a_1(c_L)} & \text{if } c = c_L \end{cases} \tag{8}$$

$$P(c|s_2 = c_L, a_1 = 1) = \begin{cases} \frac{(1-q)(qa_1(c_H)+(1-q)a_1(c_L))}{2q(1-q)a_1(c_H)+(q^2+(1-q)^2)a_1(c_L)} & \text{if } c = c_H \\ \frac{q((1-q)a_1(c_H)+qa_1(c_L))}{2q(1-q)a_1(c_H)+(q^2+(1-q)^2)a_1(c_L)} & \text{if } c = c_L \end{cases} \tag{9}$$

and

$$P(c|s_2 = c_H, a_1 = 0) = \begin{cases} \frac{q(q(1-a_1(c_H))+(1-q)(1-a_1(c_L)))}{(q^2+(1-q)^2)(1-a_1(c_H))+2q(1-q)(1-a_1(c_L))} & \text{if } c = c_H \\ \frac{(1-q)((1-q)(1-a_1(c_H))+q(1-a_1(c_L)))}{(q^2+(1-q)^2)(1-a_1(c_H))+2q(1-q)(1-a_1(c_L))} & \text{if } c = c_L \end{cases} \tag{10}$$

$$P(c|s_2 = c_L, a_1 = 0) = \begin{cases} \frac{(1-q)(q(1-a_1(c_H))+(1-q)(1-a_1(c_L)))}{2q(1-q)(1-a_1(c_H))+(q^2+(1-q)^2)(1-a_1(c_L))} & \text{if } c = c_H \\ \frac{q((1-q)(1-a_1(c_H))+q(1-a_1(c_L)))}{2q(1-q)(1-a_1(c_H))+(q^2+(1-q)^2)(1-a_1(c_L))} & \text{if } c = c_L \end{cases} \tag{11}$$

We can write the expectation as:

$$E[c|s_2, a_1] = c_L P(c_L|s_2, a_1) + c_H P(c_H|s_2, a_1)$$

and since we have defined $c_H = 1$ and $c_L = 0$ we can write:

$$E[c|s_2, a_1] = P(c_H|s_2, a_1)$$

If it is not the case that either $a_1(c_H) = a_1(c_L) = 1$ or $a_1(c_H) = a_1(c_L) = 0$ then we can write expression (2) conditional on c_L as:

$$U_1(c_L) = a_1(c_L)(u(w_1) - (1-q)) \tag{12}$$
$$+(1-q)a_1(c_L)a_2(c_H, a_1 = 1)(u(w_2) - \left(\frac{q^2 a_1(c_H) + q(1-q)a_1(c_L)}{(q^2 + (1-q)^2)a_1(c_H) + 2q(1-q)a_1(c_L)}\right))$$
$$+(1-q)(1-a_1(c_L))a_2(c_H, a_1 = 0)(u(w_2) -$$
$$\left(\frac{q^2(1-a_1(c_H)) + q(1-q)(1-a_1(c_L))}{(q^2 + (1-q)^2)(1-a_1(c_H)) + 2q(1-q)(1-a_1(c_L))}\right))$$
$$+qa_1(c_L)a_2(c_L, a_1 = 1)(u(w_2) - \left(\frac{q(1-q)a_1(c_H) + (1-q)^2 a_1(c_L)}{2q(1-q)a_1(c_H) + (q^2 + (1-q)^2)a_1(c_L)}\right))$$
$$+q(1-a_1(c_L))a_2(c_L, a_1 = 0)(u(w_2) - \left(\frac{q(1-q)(1-a_1(c_H)) + (1-q)^2(1-a_1(c_L))}{2q(1-q)(1-a_1(c_H)) + (q^2 + (1-q)^2)(1-a_1(c_L))}\right))$$

and the analogous expression for $U_1(c_H)$.

6.2 Supporting Results

Lemma 1 simplifies the decision problem as, out of the four second period actions, the value of at most one is determined by first period strategies.

Lemma 1 *Given Conditions M and O:*

$$\frac{q^2}{q^2 + (1-q)^2} \geq E_2[c|s_H, a_1 = 0] \geq q \geq E_2[c|s_H, a_1 = 1] \geq 0.5$$
$$\geq E_2[c|s_L, a_1 = 0] \geq 1 - q \geq E_2[c|s_H, a_1 = 1] \geq \frac{(1-q)^2}{q^2 + (1-q)^2}$$

Proof of Lemma 1: When it is not the case that either $a_1(c_H) = a_1(c_L) = 1$ or $a_1(c_H) = a_1(c_L) = 0$

$$E[c|s_H, a_1 = 0] = \left(\frac{q^2(1-a_1(c_H)) + q(1-q)(1-a_1(c_L))}{(q^2 + (1-q)^2)(1-a_1(c_H)) + 2q(1-q)(1-a_1(c_L))}\right)$$
$$E[c|s_H, a_1 = 1] = \left(\frac{q^2 a_1(c_H) + q(1-q)a_1(c_L)}{(q^2 + (1-q)^2)a_1(c_H) + 2q(1-q)a_1(c_L)}\right)$$
$$E[c|s_L, a_1 = 0] = \left(\frac{q(1-q)(1-a_1(c_H)) + (1-q)^2(1-a_1(c_L))}{2q(1-q)(1-a_1(c_H)) + (q^2 + (1-q)^2)(1-a_1(c_L))}\right)$$
$$E[c|s_L, a_1 = 1] = \left(\frac{q(1-q)a_1(c_H) + (1-q)^2 a_1(c_L)}{2q(1-q)a_1(c_H) + (q^2 + (1-q)^2)a_1(c_L)}\right)$$

Given condition M it must be that:

$$\begin{aligned}
\frac{q^2}{q^2+(1-q)^2} &\geq E[c|s_H, a_1 = 0] \geq q \\
q &\geq E[c|s_H, a_1 = 1] \geq 0.5 \\
0.5 &\geq E[c|s_L, a_1 = 0] \geq 1-q \\
1-q &\geq E[c|s_L, a_1 = 1] \geq \frac{(1-q)^2}{q^2+(1-q)^2}
\end{aligned}$$

If $a_1(c_H) = a_1(c_L) = 1$ then given condition O:

$$\begin{aligned}
E[c|s_H, a_1 &= 0] = \frac{q^2}{q^2+(1-q)^2} \\
E[c|s_H, a_1 &= 1] = q \\
E[c|s_L, a_1 &= 0] = 0.5 \\
E[c|s_L, a_1 &= 1] = 1-q
\end{aligned}$$

If $a_1(c_H) = a_1(c_L) = 0$ then given condition O:

$$\begin{aligned}
E[c|s_H, a_1 &= 0] = q \\
E[c|s_H, a_1 &= 1] = 0.5 \\
E[c|s_L, a_1 &= 0] = 1-q \\
E[c|s_L, a_1 &= 1] = \frac{(1-q)^2}{q^2+(1-q)^2}
\end{aligned}$$

■

Lemma 2 *If $u(w_1) > 1-q$ then*

$$a_1(c_L) = 1$$

will be an SJO best response to every $a_1(c_H) \in [0,1]$.

Proof of Lemma: Suppose that $a_1(c_H) = 1$ then $a_1(c_L) = 1$ by Condition M.

Suppose that $a_1(c_H) = 0$. In this case for $a_1(c_L) > 0$:

$$\begin{aligned}
U_1(c_L) = a_1(c_L)(u(w_1) - (1-q)) + (1-q)a_1(c_L)a_2(c_H, a_1 = 1)(u(w_2) - \frac{1}{2}) \qquad (13) \\
+(1-q)(1-a_1(c_L))a_2(c_H, a_1 = 0)(u(w_2) - \left(\frac{q^2 + q(1-q)(1-a_1(c_L))}{(q^2+(1-q)^2) + 2q(1-q)(1-a_1(c_L))}\right)) \\
+qa_1(c_L)a_2(c_L, a_1 = 1)(u(w_2) - \left(\frac{(1-q)^2}{q^2+(1-q)^2}\right)) \\
+q(1-a_1(c_L))a_2(c_L, a_1 = 0)(u(w_2) - \left(\frac{q(1-q) + (1-q)^2(1-a_1(c_L))}{2q(1-q) + (q^2+(1-q)^2)(1-a_1(c_L))}\right))
\end{aligned}$$

Since $\frac{1}{2} < \frac{q^2+q(1-q)(1-a_1(c_L))}{(q^2+(1-q)^2)+2q(1-q)(1-a_1(c_L))}$ and $\frac{(1-q)^2}{q^2+(1-q)^2} < \frac{q(1-q)+(1-q)^2(1-a_1(c_L))}{2q(1-q)+(q^2+(1-q)^2)(1-a_1(c_L))}$ for every $a_1(c_L)$, and $u(w_1) > 1-q$, $U_1(c_L)$ is an increasing function of $a_1(c_L)$ with a maximum at 1. Therefore for $a_1(c_L) > 0$, $a_1(c_L) = 1$ is a best response to $a_1(c_H) = 0$.

In the case that $a_1(c_L) = 0$:

$$U_1(c_L) = (1-q)a_2(c_H, a_1 = 0)(u(w_2) - q) + a_2(c_L, a_1 = 0)(u(w_2) - (1-q)) \tag{14}$$

Selecting $a_1(c_L) > 0$ does strictly better than $a_1(c_L) = 0$ as expression (13) is always greater than expression (14). And so if $a_1(c_H) = 0$ then $a_1(c_L) = 1$ is a best response.

Suppose that $a_1(c_H) \in (0,1)$ then expression (12) applies. Just as in the $a_1(c_H) = 0$ case $U_1(c_L)$ is an increasing function of $a_1(c_L)$ (with the domain of $a_1(c_L) \in [0,1]$) with a maximum at 1 and the Lemma is proved.

■

Lemma 3 *Suppose that $a, a', b, b', c, c' > 0$. If $1 \geq y > x \geq 0$ and $\frac{a}{a'} \geq \frac{b}{b'} \geq \frac{c}{c'}$, with at least one inequality strict then*

$$\frac{ay^2 + by + c}{a'y^2 + b'y + c'} > \frac{ax^2 + bx + c}{a'x^2 + b'x + c'} \tag{15}$$

Proof: We rewrite the expression (15) as:

$$\begin{aligned} & ab'xy^2 + ac'y^2 + a'bx^2y + bc'y + a'cx^2 + b'cx \\ > \; & a'bxy^2 + a'cy^2 + ab'x^2y + b'cy + ac'x^2 + bc'x \end{aligned}$$

which is equivalent to:

$$xy(y-x)(ab' - a'b) + (y^2 - x^2)(ac' - a'c) + (y-x)(bc' - b'c) > 0$$

the above expression always holds when $\frac{a}{a'} \geq \frac{b}{b'} \geq \frac{c}{c'}$ and when at least one inequality holds strictly. Therefore the lemma is proved.

■

Lemma 4 *If $u(w_1), u(w_2) \in (1-q, \frac{1}{2}]$ then it is SJO that $a_1^*(c_H) = 0$ and $a_1^*(c_L) = 1$.*

Proof of Lemma 4: By Lemma 2, it must be that $a_1^*(c_L) = 1$ which implies that $E[c|c_L, a_1 = 0] = \frac{1}{2}$. Therefore, by assumption $u(w_2) \leq \frac{1}{2}$ it must be that $a_2^*(c_L, a_1 = 0) = 0$. Since $u(w_2) > 1-q$ it will be that $a_2^*(c_L, a_1 = 1) = 1$. We can write

$$\begin{aligned} U_1(c_H) &= a_1(c_H)(u(w_1) - q) \\ &+(1-q)a_1(c_H)(u(w_2) - \left(\frac{q(1-q)a_1(c_H) + (1-q)^2}{2q(1-q)a_1(c_H) + (q^2 + (1-q)^2)}\right)) \end{aligned}$$

Therefore

$$\frac{\partial U_1(c_H)}{\partial a_1(c_H)} = u(w_1) - q$$
$$+(1-q)(u(w_2) - \left(\frac{2q^2(1-q)^2a_1(c_H)^2 + 2q(1-q)(q^2+(1-q)^2)a_1(c_H) + (1-q)^2(q^2+(1-q)^2)}{4q^2(1-q)^2a_1(c_H)^2 + 4q(1-q)(q^2+(1-q)^2)a_1(c_H) + (q^2+(1-q)^2)^2}\right))$$

By Lemma 3, $\frac{\partial U_1(c_H)}{\partial a_1(c_H)}$ is everywhere decreasing in $a_1(c_H)$. Additionally, $\frac{\partial U_1(c_H)}{\partial a_1(c_H)}$ is everywhere strictly less than zero as

$$\frac{1}{2} - q + (1-q)\left(\frac{1}{2} - \frac{(1-q)^2}{q^2+(1-q)^2}\right) = \frac{1}{2} - q + (1-q)\left(\frac{q-\frac{1}{2}}{q^2+(1-q)^2}\right)$$

which is less than zero and so it must be that $a_1^*(c_H) = 0$.

■

Lemma 5 *For every $u(w_2) \in (\frac{1}{2}, q)$ there exists a $\lambda^*(u(w_2)) \in [\frac{1}{2}, q)$ such that for all $u(w_1) \in (\frac{1}{2}, \lambda^*(u(w_2)))$ it is SJO that $a_1^*(c_H) = 0$ and $a_1^*(c_L) = 1$ and for all $u(w_1) \in (\lambda^*(u(w_2)), q)$ it is SJO that $a_1^*(c_H) \in (0,1)$ and $a_1^*(c_L) = 1$. Additionally, if $u(w_1) = u(w_2)$ then $\lambda^* > \frac{1}{2}$ and $a_2^*(c_H, a_1 = 1) = 1$*

Proof: By Lemma 2 it must be that $a_1^*(c_L) = 1$. Since $u(w_1), u(w_2) \in (\frac{1}{2}, q)$ it will be that

$$a_2^*(c_L, a_1 = 0) = a_2^*(c_L, a_1 = 1) = 1$$
$$a_2^*(c_H, a_1 = 0) = 0$$

We can write:

$$U_1(c_H) = a_1(c_H)(u(w_1) - q) \tag{16}$$
$$+(1-q)a_1(c_H)(u(w_2) - \left(\frac{q(1-q)a_1(c_H) + (1-q)^2}{2q(1-q)a_1(c_H) + (q^2+(1-q)^2)}\right))$$
$$+(1-q)(1-a_1(c_H))(u(w_2) - \frac{1}{2})$$
$$+qa_1(c_H)a_2(c_H, a_1 = 1)(u(w_2) - \left(\frac{q^2a_1(c_H) + q(1-q)}{(q^2+(1-q)^2)a_1(c_H) + 2q(1-q)}\right))$$

and therefore

$$\frac{\partial U_1(c_H)}{\partial a_1(c_H)} = u(w_1) - q \tag{17}$$
$$+(1-q)(0.5 - \left[\frac{2q^2(1-q)^2a_1(c_H)^2 + 2q(1-q)(q^2+(1-q)^2)a_1(c_H) + (1-q)^2(q^2+(1-q)^2)}{4q^2(1-q)^2a_1(c_H)^2 + 4q(1-q)(q^2+(1-q)^2)a_1(c_H) + (q^2+(1-q)^2)^2}\right])$$
$$+qa_2(c_H, a_1 = 1)(u(w_2) - \left[\frac{q^2(q^2+(1-q)^2)a_1(c_H)^2 + 4q^3(1-q)a_1(c_H) + 2q^2(1-q)^2}{(q^2+(1-q)^2)^2a_1(c_H)^2 + 4q(1-q)a_1(c_H) + 4q^2(1-q)^2}\right])$$

By Lemma 3, both of the terms in brackets in expression (17) are strictly increasing in $a_1(c_H)$. Therefore, $\frac{\partial U_1(c_H)}{\partial a_1(c_H)}$ is everywhere strictly decreasing and for either value of $a_2(c_H, a_1 = 1)$ there is at most one value of $a_1(c_H)$ such that $\frac{\partial U_1(c_H)}{\partial a_1(c_H)} = 0$. We write:

$$\left.\frac{\partial U_1(c_H)}{\partial a_1(c_H)}\right|_{a_1(c_H)=0} = u(w_1) - q + (1-q)(\frac{1}{2} - \left(\frac{(1-q)^2}{q^2 + (1-q)^2}\right)) + qa_2(c_H, a_1 = 1)(u(w_2) - \frac{1}{2})$$

We define:

$$\lambda(u(w_2)) = q - \frac{(1-q)(q-\frac{1}{2})}{q^2 + (1-q)^2} - qa_2(c_H, a_1 = 1)(u(w_2) - \frac{1}{2})$$

Note that $\lambda(u(w_2)) < q$ will always hold. However, for large $u(w_2)$ it can be that $\lambda(u(w_2)) < \frac{1}{2}$. Therefore if $\lambda(u(w_2)) \in [\frac{1}{2}, q)$ then we define $\lambda(u(w_2)) = \lambda^*(u(w_2))$. If $\lambda(u(w_2)) < \frac{1}{2}$ then we define $\lambda^*(u(w_2)) = \frac{1}{2}$. Therefore for $u(w_2) > \lambda^*$ it must be that $a_1^*(c_H) \in (0, 1)$.

Note that $a_2^*(c_H, a_1 = 1) = 1$ if and only if:

$$u(w_2) > \frac{q^2 a_1(c_H) + q(1-q)}{(q^2 + (1-q)^2)a_1(c_H) + 2q(1-q)}$$

which implies

$$a_1(c_H)[u(w_2)(q^2 + (1-q)^2) - q^2] > q(1-q) - u(w_2)(2q(1-q))$$

As $u(w_2)(q^2 + (1-q)^2) - q^2$ is negative we write

$$a_1(c_H) < \frac{q(1-q) - u(w_2)(2q(1-q))}{u(w_2)(q^2 + (1-q)^2) - q^2} \tag{18}$$

We define the right side of expression (18) as $\widehat{a}$ such that:

$$\begin{aligned} a_1(c_H) &< \widehat{a} \text{ if and only if } a_2^*(c_H, a_1 = 1) = 1 \\ a_1(c_H) &\geq \widehat{a} \text{ if and only if } a_2^*(c_H, a_1 = 1) = 0 \end{aligned}$$

Observe that $\widehat{a} = 0$ when $u(w_2) = \frac{1}{2}$, $\widehat{a} = 1$ when $u(w_2)$ and that $\widehat{a}$ is strictly increasing on this region, as $\frac{\partial \widehat{a}}{\partial u(w_2)} = \frac{q(1-q)(2q-1)}{(q^2 - u(w_2)(2q^2 - 2q + 1))^2} > 0$.

Now consider that $u(w_1) = u(w_2)$. If $a_2^*(c_H, a_1 = 1) = 1$ then

$$u(w_1) = u(w_2) \leq \frac{q - \frac{(1-q)(q-\frac{1}{2})}{q^2 + (1-q)^2} + \frac{1}{2}q}{1+q} = \underline{\lambda}$$

is equivalent to $a_1^*(c_H) = 0$. If $a_2^*(c_H, a_1 = 1) = 0$ then

$$u(w_1) = u(w_2) \leq q - \frac{(1-q)(q-\frac{1}{2})}{q^2 + (1-q)^2} = \overline{\lambda}$$

is equivalent to $a_1^*(c_H) = 0$ where $\underline{\lambda} < \overline{\lambda}$. By expression (16) it will then be optimal for $a_1^*(c_H)$ to be determined where $a_2^*(c_H, a_1 = 1) = 1$.

■

Lemma 6 *For every* $u(w_2) \in (q, \frac{q^2}{q^1+(1-q)^2})$ *there is a* $\mu^*(u(w_2)) \in [q, \frac{q^2}{q^1+(1-q)^2}]$ *such that for all* $u(w_1) \in (q, \mu^*(u(w_2)))$ *then it is SJO that* $a_1^*(c_H) \in (0,1)$ *and* $a_1^*(c_L) = 1$ *and for all* $u(w_1) \in (\mu^*(u(w_2)), \frac{q^2}{q^2+(1-q)^2})$ *then it is SJO that* $a_1^*(c_H) = a_1^*(c_L) = 1$.

Proof of Lemma 6: By Lemma 2 it must be that $a_1^*(c_L) = 1$. This implies that $E[c|c_H, a_1 = 0] = \frac{q^2}{q^2+(1-q)^2}$ therefore $a_2^*(c_H, a_1 = 0) = 0$. We write the utility as:

$$\begin{aligned} U_1(c_H) &= a_1(c_H)(u(w_1) - q) \\ &+(1-q)a_1(c_H)(u(w_2) - \left(\frac{q(1-q)a_1(c_H) + (1-q)^2}{2q(1-q)a_1(c_H) + (q^2 + (1-q)^2)}\right)) \\ &+(1-q)(1-a_1(c_H))(u(w_2) - \frac{1}{2}) \\ &+qa_1(c_H)(u(w_2) - \left(\frac{q^2 a_1(c_H) + q(1-q)}{(q^2 + (1-q)^2)a_1(c_H) + 2q(1-q)}\right)) \end{aligned}$$

Since this matches expression (16) in the proof of Lemma 5 with the exception that $a_2^*(c_H, a_1 = 1) = 1$, much of the reasoning, without the complications of determining $a_2^*(c_H, a_1 = 1)$, carries over. Expression (17) is also valid here:

$$\begin{aligned} \frac{\partial U_1(c_H)}{\partial a_1(c_H)} &= u(w_1) - q \\ +(1-q)(0.5 - &\left[\frac{2q^2(1-q)^2 a_1(c_H)^2 + 2q(1-q)(q^2 + (1-q)^2)a_1(c_H) + (1-q)^2(q^2 + (1-q)^2)}{4q^2(1-q)^2 a_1(c_H)^2 + 4q(1-q)(q^2 + (1-q)^2)a_1(c_H) + (q^2 + (1-q)^2)^2}\right]) \\ +qa_1(c_H)(u(w_2) - &\left[\frac{q^2(q^2 + (1-q)^2)a_1(c_H)^2 + 4q^3(1-q)a_1(c_H) + 2q^2(1-q)^2}{(q^2 + (1-q)^2)^2 a_1(c_H)^2 + 4q(1-q)a_1(c_H) + 4q^2(1-q)^2}\right]) \end{aligned}$$

Again, $\frac{\partial U_1(c_H)}{\partial a_1(c_H)}$ is strictly decreasing and equals zero at most once. It can never be that $a_1^*(c_H) = 0$ because

$$\left.\frac{\partial U_1(c_H)}{\partial a_1(c_H)}\right|_{a_1(c_H)=0} = (1-q)\left(0.5 - \frac{(1-q)^2}{q^2 + (1-q)^2}\right) > 0$$

Therefore,

$$\left.\frac{\partial U_1(c_H)}{\partial a_1(c_H)}\right|_{a_1(c_H)=1} = u(w_1) - q + (1-q)(0.5 + 2q^3 - 3q^2 + 2q - 1) \tag{19}$$
$$+qu(w_2) - \left(\frac{q^2(3-2q)}{8q^2(1-q)^2+1}\right)$$

And so we define:

$$\mu(u(w_2)) = \left(\frac{q + (1-q)(0.5 - 2q^3 + 3q^2 - 2q)}{1+q}\right) + \left(\frac{q^2(3-2q)}{8q^2(1-q)^2+1}\right) \tag{20}$$

For $\mu(u(w_2)) \in [q, \frac{q^2}{q^2+(1-q)^2}]$ we set $\mu(u(w_2)) = \mu^*(u(w_2))$. If $\mu(u(w_2)) < q$ (because $u(w_2)$ is in the high end of the range) then we set $\mu^*(u(w_2)) = q$. In this case $a_1^*(c_H) = 1$ is SJO for every $u(w_1) \in (q, \frac{q^2}{q^2+(1-q)^2})$. If $\mu(u(w_2)) > \frac{q^2}{q^2+(1-q)^2}$ (because $u(w_2)$ is in the low end of the range) then we set $\mu^*(u(w_2)) = \frac{q^2}{q^2+(1-q)^2}$. In this case $a_1^*(c_H) \in (0,1)$ is SJO for every $u(w_1) \in (q, \frac{q^2}{q^2+(1-q)^2})$.

From expression (20) note that for every value of $q \in (0.5, 1)$ it will be that

$$\frac{q^2}{q^2+(1-q)^2} > \mu(u(w_2))$$

However, for some values of q,

$$\mu(u(w_2)) > q$$

is violated.

■

Lemma 7 *If $u(w_1)$ and $u(w_2) > \frac{q^2}{q^2+(1-q)^2}$ then it is SJO that $a_1^*(c_H) = a_1^*(c_L) = 1$.*

Proof of Lemma 7: By Lemma 2 it must be that $a_1^*(c_L) = 1$. This implies that $E[c|c_H, a_1 = 0] = \frac{q^2}{q^2+(1-q)^2}$ therefore $a_2(c_H, a_1 = 0) = 1$. We write the utility as:

$$U_1(c_H) = a_1(c_H)(u(w_1) - q)$$
$$+(1-q)a_1(c_H)(u(w_2) - \left(\frac{q(1-q)a_1(c_H) + (1-q)^2}{2q(1-q)a_1(c_H) + (q^2+(1-q)^2)}\right))$$
$$+(1-q)(1-a_1(c_H))(u(w_2) - 0.5)$$
$$+qa_1(c_H)(u(w_2) - \left(\frac{q^2a_1(c_H) + q(1-q)}{(q^2+(1-q)^2)a_1(c_H) + 2q(1-q)}\right))$$
$$+q(1-a_1(c_H))(u(w_2) - \left(\frac{q^2}{(q^2+(1-q)^2)}\right))$$

Therefore

$$\frac{\partial U_1(c_H)}{\partial a_1(c_H)} = u(w_1) - q$$
$$+(1-q)(0.5 - \left[\frac{2q^2(1-q)^2a_1(c_H)^2 + 2q(1-q)(q^2+(1-q)^2)a_1(c_H) + (1-q)^2(q^2+(1-q)^2)}{4q^2(1-q)^2a_1(c_H)^2 + 4q(1-q)(q^2+(1-q)^2)a_1(c_H) + (q^2+(1-q)^2)^2}\right])$$
$$+q(\frac{q^2}{q^2+(1-q)^2} - \left[\frac{q^2(q^2+(1-q)^2)a_1(c_H)^2 + 4q^3(1-q)a_1(c_H) + 2q^2(1-q)^2}{(q^2+(1-q)^2)^2a_1(c_H)^2 + 4q(1-q)a_1(c_H) + 4q^2(1-q)^2}\right])$$

Again by Lemma 3 the function $\frac{\partial U_1(c_H)}{\partial a_1(c_H)}$ is strictly decreasing. It follows from the proof of Lemma 6 that for all $a_1(c_H)$:

$$\frac{\partial U_1(c_H)}{\partial a_1(c_H)} > 0$$

and so it must be that $a_1^*(c_H) = 1$.

■

6.3 Proofs of Main Results

Proof of Proposition 1: By way of contradiction, suppose that $a_1^*(c_H) = a_1^*(c_L) = z \in (0,1)$. This implies that

$$\begin{aligned} E[c|c_H, a_1 &= 1] = E[c|c_H, a_1 = 0] = E[c|c_H] = q \\ E[c|c_L, a_1 &= 1] = E[c|c_L, a_1 = 0] = E[c|c_L] = 1-q \end{aligned}$$

and so:

$$\begin{aligned} a_2^*(c_H, a_1 &= 0) = a_2^*(c_H, a_1 = 1) = a_2^*(c_H) \\ a_2^*(c_L, a_1 &= 0) = a_2^*(c_L, a_1 = 1) = a_2^*(c_L) \end{aligned}$$

For SJO:

$$U_1(c_L, a_1(c_H) = z) \geq U_1(c_L, a_1(c_H) = z') \text{ for all other } z' \text{ and} \quad (21)$$
$$U_1(c_{H,}a_1(c_L) = z) \geq U_1(c_{H,}a_1(c_L) = z') \text{ for all other } z' \quad (22)$$

yielding

$$\begin{aligned} U_1^*(c_L) &= z(u(w_1) - (1-q)) + qa_2^*(c_L)[u(w_2) - (1-q)] + (1-q)a_2^*(c_H)[u(w_2) - q] \\ U_1^*(c_H) &= z(u(w_1) - q) + qa_2^*(c_H)[u(w_2) - q] + (1-q)a_2^*(c_L)[u(w_2) - (1-q)] \end{aligned}$$

If $u(w_1) - (1-q) > 0$ then increasing z increases both quantities in and therefore expressions (21) and (22) cannot hold. If $u(w_1) - q < 0$ then reducing z will increase both values of U_1 and therefore expressions (21) and (22) cannot hold. If $u(w_1) - (1-q) > 0 > u(w_1) - q$ then according to Lemma 2 there exists a profitable deviation to $a_1(c_L) = 1$ and therefore

expression (21) cannot hold. Therefore the strategy cannot be SJO.

■

Proof of Proposition 5: With a perfect memory agent the second period posteriors and expectations are standard:

$$\begin{aligned} E[c|s_1 &= c_H, s_2 = c_H] = \frac{q^2}{q^2+(1-q)^2} \\ E[c|s_1 &= c_H, s_2 = c_L] = \frac{1}{2} \\ E[c|s_1 &= c_L, s_2 = c_H] = \frac{1}{2} \\ E[c|s_1 &= c_L, s_2 = c_L] = \frac{(1-q)^2}{q^2+(1-q)^2} \end{aligned}$$

as first period actions do not affect second period beliefs. The ex-ante utility of the agent is:

$$\begin{aligned} U_0(w) = \frac{1}{2}\{a_1(c_L)[u(w_1)-(1-q)] \\ +qa_2(c_L,c_L)[u(w_2)-\frac{(1-q)^2}{q^2+(1-q)^2}]+(1-q)a_2(c_L,c_H)[u(w_2)-\frac{1}{2}]]\} \\ +\frac{1}{2}\{a_1(c_H)[u(w_1)-q] \\ +qa_2(c_H,c_H)[u(w_2)-\frac{q^2}{q^2+(1-q)^2}]+(1-q)a_2(c_H,c_L)[u(w_2)-\frac{1}{2}]\} \end{aligned}$$

From this it follows that the $a_1^*(c_L)$ and $a_1^*(c_H)$ are selected as if the agent was myopic. Suppose that Definitions 2 and 3 are satisfied. Together with the observation of myopic behavior, closeness implies that $a_1^{*C}(c_H) = a_1^{*I}(c_H)$ and $a_1^{*C}(c_L) = a_1^{*I}(c_L)$. Also from closeness it must be that:

$$a_2^{*C}(s_1,s_2) = a_2^{*I}(s_1,s_2) \text{ for } s_1,s_2 \in \{c_L,c_H\}$$

Therefore expression (4) in Definition 2:

$$\begin{aligned} & u(w_1^C)\Psi_1^C + u(w_2^C)\Psi_2^C \\ > \; & u(w_1^I)\Psi_1^I + u(w_2^I)\Psi_2^I \end{aligned}$$

is equivalent to

$$U_0(w^C) - U_0(w^I) > 0$$

Therefore the agent cannot display PIP.

■

Proof of Proposition 6: Because first period and second period behavior are identical

in both contracts we can write:

$$U_0(w^C) - U_0(w^I) = \left(\frac{a_1(c_L) + a_1(c_H)}{2}\right)(u(w_1^C) - u(w_1^I))$$
$$+\frac{1}{2}\{(qa_1(c_H) + (1-q)a_1(c_L))a_2(c_H, a_1 = 1)$$
$$+(q(1-a_1(c_H)) + (1-q)(1-a_1(c_L)))a_2(c_H, a_1 = 0)$$
$$+((1-q)a_1(c_H) + qa_1(c_L))a_2(c_L, a_1 = 1)$$
$$+((1-q)(1-a_1(c_H)) + q(1-a_1(c_L)))a_2(c_L, a_1 = 0)\}(u(w_2^C) - u(w_2^I))$$

This can be rewritten as expression (4) in Definition 2 and so it must be that

$$U_0(w^C) - U_0(w^I) > 0$$

and so PIP can never occur.

■

Proof of Proposition 2: If $u(w_1), u(w_2) > \frac{q^2}{q^1+(1-q)^2}$ then Lemma 7 shows that $a_1^*(c_L) = a_1^*(c_H) = 1$. If $u(w_1), u(w_2) \in (1-q, \frac{1}{2}]$ then Lemma 4 shows that $a_1^*(c_L) = 1$ and $a_1^*(c_H) = 0$. Under both cases Proposition 6 applies and so Proposition 2 is proved.

■

Proof of Proposition 3: Lemma 5 shows that such a λ^* exists. Further Lemma 5 shows that for $u(w_1), u(w_2) \in (\frac{1}{2}, \lambda^*)$ that $a_1^*(c_L) = 1$ and $a_1^*(c_H) = 0$. Lemma 6 shows that such a μ^* exists and that for $u(w_1), u(w_2) \in (\mu^*, \frac{q^2}{q^1+(1-q)^2})$ that $a_1^*(c_L) = a_1^*(c_H) = 1$. Under both cases Proposition 6 applies and so Proposition 3 is proved.

■

The following Lemmas will be helpful in the Proof of Proposition 4.

Lemma 8 *Suppose that the SJO is such that* $a_1^*(c_L) = 1$ *and* $a_1^*(c_H) \in (0,1)$ *for* w *and* w'. *If*

$$u(w_1) + qu(w_2) > u(w_1') + qu(w_2')$$

then $a_1^*(c_H) > a_1'^*(c_H)$

Proof: By assumption, $a_1^*(c_H)$ has an interior maximum which is determined by:

$$\left.\frac{\partial U_1(c_H)}{a_1(c_H)}\right|_{a_1^*(c_H)} = 0$$

where

$$\frac{\partial U_1(c_H)}{a_1(c_H)} = u(w_1) - q$$
$$+(1-q)(0.5 - \left[\frac{2q^2(1-q)^2a_1(c_H)^2 + (q^2+(1-q)^2)(2q(1-q)a_1(c_H) + (1-q)^2)}{(2q(1-q)a_1(c_H) + q^2 + (1-q)^2)^2}\right])$$
$$+q(u(w_2) - \left[\frac{q^2(q^2+(1-q)^2)a_1(c_H)^2 + (2q(1-q))(2q^2a_1(c_H) + q(1-q))}{((q^2+(1-q)^2)a_1(c_H) + 2q(1-q))^2}\right])$$

By Lemma 3 the expressions in the brackets are strictly increasing on $a_1(c_H) \in [0,1]$ for every $q \in (\frac{1}{2}, 1)$. Since

$$u(w_1) - u(w_1') + q(u(w_2) - u(w_2')) > 0$$

it follows that $a_1^*(c_H) > a_1'^*(c_H)$.

■

Lemma 9 *If $a_1^*(c_L) = 1$ and $a_1^*(c_H) \in (0,1)$ then for every $\widehat{a}_1(c_H)$ it will be that*

$$\frac{\partial U_1(c_H)}{\partial a_1(c_H)}\Big|_{\widehat{a}_1(c_H)} - \frac{\partial U_0}{\partial a_1(c_H)}\Big|_{\widehat{a}_1(c_H)} = q(1-q)(2q-1) > 0$$

Proof: For these parameter values we can write:

$$\frac{\partial U_1(c_H)}{a_1(c_H)} = u(w_1) - q$$
$$+(1-q)(0.5 - \left[\frac{2q^2(1-q)^2a_1(c_H)^2 + 2q(1-q)(q^2+(1-q)^2)a_1(c_H) + (1-q)^2(q^2+(1-q)^2)}{4q^2(1-q)^2a_1(c_H)^2 + 4q(1-q)(q^2+(1-q)^2)a_1(c_H) + (q^2+(1-q)^2)^2}\right])$$
$$+q(u(w_2) - \left[\frac{q^2(q^2+(1-q)^2)a_1(c_H)^2 + 4q^3(1-q)a_1(c_H) + 2q^2(1-q)^2}{(q^2+(1-q)^2)^2 + 4q(1-q)(q^2+(1-q)^2)a_1(c_H) + 4q^2(1-q)^2}\right])$$

and

$$\frac{\partial U_0}{\partial a_1(c_H)} = u(w_1) - q$$
$$+(1-q)(0.5 - \left(\frac{2q^2(1-q)^2a_1(c_H)^2 + 2q(1-q)(q^2+(1-q)^2)a_1(c_H) + q^4 + (1-q)^4}{4q^2(1-q)^2a_1(c_H)^2 + 4q(1-q)(q^2+(1-q)^2)a_1(c_H) + (q^2+(1-q)^2)^2}\right))$$
$$+q(u(w_2) - \left(\frac{q^2(q^2+(1-q)^2)a_1(c_H)^2 + 4q^3(1-q)a_1(c_H) + (1-q)^2(2q^2+2q-1)}{(q^2+(1-q)^2)^2 + 4q(1-q)(q^2+(1-q)^2)a_1(c_H) + 4q^2(1-q)^2}\right))$$

Mercifully the difference simplifies to:

$$\frac{\partial U_1(c_H)}{\partial a_1(c_H)} - \frac{\partial U_0}{\partial a_1(c_H)} = (1-q)(-q^2 + 2q^3) + q(1-q)^2(2q-1)$$

So the Lemma is proved.

■

Proof of Proposition 4: We offer a constructive proof where we identify the process of determining a contract w^I sufficient to satisfy PIP. We proceed in two main parts. Part 1 shows that a contract w^I which induces first period actions so that $a_1^C > a_1^I$ will imply that

$$2(U_0(w^C) - U_0(w^I)) - \varepsilon < 0$$

Part 2 shows that a contract w^I can be found which makes ε arbitrarily small.

Part 1:

For the proof to follow it must be that $a_1^*(c_H) \in (0,1)$ and $a_1^*(c_L) = 1$. This is the content of Lemmas 5 and 6. In the exhibition of PIP it is required that given a constant contract w^C we can find an increasing contract w^I preferred by the ex-ante player but satisfying the requirements provided in the definition. Given the increasing contract w^C,we select the increasing contract w^I such that the former is more valuable:

$$\begin{aligned}&\Psi_1^C u(w_1^C) + \Psi_2^C u(w_2^C)\\ &= \Psi_1^I u(w_1^I) + \Psi_2^I u(w_2^I) + \varepsilon\end{aligned}$$

We can then rewrite the difference in ex-ante utilities as

$$\begin{aligned}&2(U_0(w^C) - U_0(w^I)) - \varepsilon \\ &= a_1^I(c_H)q - a_1^C(c_H)q \\ &+((1-q)a_1^I(c_H)+q)\left(\frac{q(1-q)a_1^I(c_H)+(1-q)^2}{2q(1-q)a_1^I(c_H)+(q^2+(1-q)^2)}\right) \\ &-(1-q)a_1^C(c_H)+q)\left(\frac{q(1-q)a_1^C(c_H)+(1-q)^2}{2q(1-q)a_1^C(c_H)+(q^2+(1-q)^2)}\right) \\ &+(1-q)\frac{1}{2}(a_1^C(c_H)-a_1^I(c_H)) \\ &+(qa_1^I(c_H)+(1-q))\left(\frac{q^2a_1^I(c_H)+q(1-q)}{(q^2+(1-q)^2)a_1^I(c_H)+2q(1-q)}\right) \\ &-(qa_1^C(c_H)+(1-q))\left(\frac{q^2a_1^C(c_H)+q(1-q)}{(q^2+(1-q)^2)a_1^C(c_H)+2q(1-q)}\right)\end{aligned} \tag{23}$$

To see that $\frac{\partial U_0}{\partial a_1(c_H)}$ is monotonically decreasing, note (as we have established) that $\frac{\partial U_1}{\partial a_1(c_H)}$ is monotonically decreasing. By Lemma 9 $\frac{\partial U_1}{\partial a_1(c_H)} - \frac{\partial U_0}{\partial a_1(c_H)}$ equals a positive constant. Therefore the right side of (23) is negative for any $a_1^C(c_H) > a_1^I(c_H)$ and so the balance of the proof consists in finding an w^I which induces an appropriately small ε.

Part 2:

We have a limited amount of room in which to find w^I. For (i) the contract w^I must be such that $u(w_1^I), u(w_2^I) \in (\lambda^*, q)$. For (ii) the contract w^I it must be that $q < \mu^*$ and $u(w_1^I), u(w_2^I) \in (q, \mu^*)$. For convenience, we denote the lower element of these sets as $\underline{u}$ and the upper element as $\overline{u}$.

Pick an preliminary $u(w_1^I)^0 = u(w_2^I)^0 = u(w^I)^0$ such that

$$u(w^I)^0 < u(w^C)$$

and

$$\begin{aligned} & (\Psi_1^C + \Psi_2^C)u(w_2^C) \\ = \; & (\Psi_1^{I0} + \Psi_2^{I0})u(w^I)^0 + \delta \end{aligned}$$

We may pick δ as small as we like because $(\Psi_1^{I0} + \Psi_2^{I0})u(w^I)^0$ is an increasing function of $u(w^I)^0$. Furthermore, by Lemma 8 any such $u(w^I)^0$ induces $a_1^C(c_H) > a_1^I(c_H)$. We can select a new $u(w_1^I)$ and $u(w_2^I)$ without affecting $a_1^I(c_H)$ (therefore not affecting the right hand side of expression (23)) if

$$u(w_1^I) + qu(w_2^I) = (1+q)u(w^I)^0$$

Define

$$f(x) = \Psi_1^I(u(w^I)^0 - qx) + \Psi_2^I(u(w_2^I)^0 + x)$$

where $x \geq 0$. As noted above we have a limited amount of room with which to work. In order to stay in the allowed region it must be that $x < x^*$ where

$$x^* = \min(\overline{u} - u(w^I)^0, q(u(w^I)^0 - \underline{u}))$$

So we require that $x \in [0, x^*)$.

$$\frac{\partial f}{\partial x} = \frac{1}{2} + q + \frac{a_1^I(c_H)}{2}$$

So for every $a_1^I(c_H)$, this value is at least $\frac{1}{2} + q$. Given x^* and q we can pick $u(w^I)^0$ such that

$$(\frac{1}{2} + q)x^* > \delta$$

This implies that

$$f(x^*) > \Psi_1^C u(w_1^C) + \Psi_2^C u(w_2^C)$$

for x^*. Therefore there are values of $x \in (0, x^*)$ such that ε in expression (23) can be made arbitrarily small. Therefore the Proposition is proved.

■

7 References

Akerlof, George and Dickens, William (1982): "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, 72(3), 307-319.

Bem, Daryl (1972): "Self-Perception Theory," in *Advances in Experimental Social Psychology*, Berkowitz, L. (Ed.), New York, Academic Press, 1-62.

Benabou, Roland and Tirole Jean (2003): "Intrinsic and Extrinsic Motivation," *Review of Economic Studies*, 70, 489-520.

Benabou, Roland and Tirole, Jean (2004): "Willpower and Personal Rules," *Journal of Political Economy*, 112 (4), 848-886.

Bernheim, Douglas and Thomadsen, Raphael (2005): "Memory and Anticipation," *Economic Journal*, 115, 271-304.

Bradley, Darren (2003): "Sleeping Beauty: a Note on Dorr's Argument for 1/3," *Analysis*, 63(3), 266-268.

Cooper, Joel (1999): "Unwanted Consequences of the Self: In search of the Motivation for Dissonance Reduction," In E. Harmon-Jones and J. Mills (Eds.) *Cognitive Dissonance: Progress on a pivotal theory in social psychology.* Washington DC: APA, 149-173.

Dorr, Cian (2002): "Sleeping Beauty: in Defense of Elga," *Analysis*, 62(4), 292-296.

Elga, Adam (2000): "Self-locating Belief and the Sleeping Beauty Problem," *Analysis*, 60(2), 143-147.

Epstein, Larry and Kopylov, Igor (2006): "Cognitive Dissonance and Choice," unpublished University of Rochester and UC Irvine.

Eyster, Erik (2002): "Rationalizing the Past: A Taste for Consistency," unpublished Oxford University.

Gigliotti, Gary and Sopher, Barry (1997): "Violations in Present-Value Maximization in Income Choice," *Theory and Decision*, 43, 45-69.

Harmon-Jones, Eddie and Mills, Judson (Eds.) (1999): *Cognitive Dissonance: Progress on a pivotal theory in social psychology.* Washington DC, APA.

Hirshleifer, David and Welch, Ivo (2002): "An Economic Approach to the Psychology of Change: Amnesia, Inertia and Impulsiveness," *Journal of Economics and Management Strategy*, 11(3), 379-421.

Konow, James (2000): "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," *American Economic Review*, 90(4), 1072-1091.

Lewis, David (2001): "Sleeping Beauty: Reply to Elga," *Analysis*, 61(3), 171-176.

Loewenstein, George and Sicherman, Nachum (1991): "Do Workers Prefer Increasing Wage Profiles?" *Journal of Labor Economics*, 9, 67-84.

Matsumoto, Dawn, Peecher, Mark and Rich, Jay (2000): "Evaluations of Outcome Sequences," *Organizational Behavior and Human Decision Processes*, 83(2), 331-352.

Monton, Bradley (2002): "Sleeping Beauty and the Forgetful Bayesian," *Analysis*, 62(1), 47-53.

Mullainathan, Sendhil (2002): "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 117 (3), 735-774.

Oxoby, Robert (2003): "Attitudes and allocations: status, cognitive dissonance, and the manipulation of attitudes," *Journal of Economic Behavior and Organization*, 52, 365-385.

Oxoby, Robert (2004): "Cognitive Dissonance, Status and Growth of the Underclass," *Economic Journal*, 114, 727-749.

Piccione, Michele and Rubenstein, Ariel (1997): "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior*, 20, 3-24.

Smith, John (2007): "Cognitive Dissonance and the Overtaking Anomaly: Psychology in the Principal-Agent Relationship," unpublished Rutgers University-Camden.

Swank, Otto (2006): "The Self-Perception Theory Versus a Dynamic Learning Model," unpublished Erasmus University Rotterdam.

Weintraub, Ruth (2004): "Sleeping Beauty: a Simple Solution," *Analysis*, 64(4), 8-10.

Wilson, Andrea (2004): "Bounded Memory and Biases in Information Processes," unpublished Harvard.

Yariv, Leeat (2006): "I'll See It When I Believe It- A Simple Model of Cognitive Consistency," unpublished California Institute of Technology.

Zanna, M, and Cooper, J. (1974): "Dissonance and the Pill: An Attribution Approach to Studying the Arousal Properties of Dissonance," *Journal of Personality and Social Psychology*, 29, 703-709.