ISSN 1988-088X

Departamento de Fundamentos del Análisis Económico II

Department of Foundations of Economic Analysis II
University of the Basque Country UPV/EHU
Avda. Lehendakari Aguirre 83
48015 Bilbao (SPAIN)
http://www.dfaeii.ehu.es

DFAE-II WP Series

2012-09

Pablo Brañas-Garza, Marisa Bucheli, María Paz Espinosa, Teresa García-Muñoz

Moral cleansing and moral licenses: experimental evidence

Moral cleansing and moral licenses: experimental evidence*

Pablo Brañas-Garza, GLOBE: Universidad de Granada, Spain

Marisa Bucheli, Universidad de la República, Uruguay

María Paz Espinosa, Universidad del País Vasco, BRIDGE, Spain

Teresa García-Muñoz, GLOBE: Universidad de Granada, Spain

ABSTRACT- Research on moral cleansing and moral self-licensing has introduced dynamic considerations in the theory of moral behavior. Past bad actions trigger negative feelings that make people more likely to engage in future moral behavior to offset them. Symmetrically, past good deeds favor a positive self-perception that creates licensing effects, leading people to engage in behavior that is less likely to be moral. In short, a deviation from a “normal state of being” is balanced with a subsequent action that compensates the prior behavior. We model the decision of an individual trying to reach the optimal level of moral self-worth over time and show that under certain conditions the optimal sequence of actions follows a regular pattern which combines good and bad actions. We conduct an economic experiment where subjects play a sequence of giving decisions (dictator games) to explore this phenomenon. We find that donation in the previous period affects present decisions and the sign is negative: participants’ behavior in every round is negatively correlated to what they did in the past. Hence donations over time seem to be the result of a regular pattern of self-regulation: moral licensing (being selfish after altruist) and cleansing (altruistic after selfish).

* We thank Pier Paolo Battigalli for his comments on a previous version of the paper. Financial aid from MICINN (ECO2009-09120 and ECO2010-17049), Gobierno Vasco (DEUI, IT-313-07) and Junta de Andalucía-Excelencia (P07.SEJ.02547) is gratefully acknowledged.

I. INTRODUCTION

How and *why* moral behavior emerges is a critical question. Moral behavior is not costless. Every single altruistic action generates a cost for the donor and thus good deeds need to come with a benefit to compensate the cost. Despite a number of classical evolutionary arguments such as kin selection –Hamilton rule- or reciprocal altruism (Fehr and Fischbauer, 2003), a series of papers have dealt with more self-centered arguments such as identity, guilt-aversion or warm-glow, that describe the benefits of being moral (see Akerlof and Kranton, 2000, Charness and Dufwenberg, 2006, Battigalli and Dufwenberg, 2007 and Aguiar et al., 2010). In this paper, we are interested in the moral self-licensing and moral cleansing literature that explore the relationship between past, present and future moral behavior.

One motivation for good deeds is their positive effect on moral self-worth. When past actions make people feel confident about their moral behavior, their moral self-regard could be high enough to allow them to engage in morally dubious behavior in the present (Zhong and Liljenquist, 2006; Merritt, Effron and Monin, 2010). This is the central argument of the moral self-licensing literature. In a review of the evidence, Merritt *et al* (2010) present the two most frequent moral-licensing mechanisms used in the literature: the moral credits and the credentials models. The moral credits model uses a moral bank account metaphor: good deeds purchase "moral credits" that diminish the discomfort of engaging in bad deeds in the future. In the credentials model, actions affect the *meaning* of future actions: the value of an ambiguous behavior will be valued through the lens of past good deeds. As a consequence, a good action gives self-license to future transgressions. Note that according to the mechanism of the first model, the licensed person gets involved in what he considers a bad action but this is not the case in the second model. So the damage to self-value is different and we may expect self-license to lead to a lower number of transgressions under the moral credits than the credentials mechanism.

In turn, immoral behavior has a negative effect on moral self-worth. After engaging in bad deeds, people follow a moral behavior to recover the lost self-worth; this mechanism is the so-called moral cleansing behavior (see Sachdeva, Iliev and Medin, 2009). One well documented example is that in response to sins, many religious practices require bodily purification.

Taking into account the two types of behavior, moral licensing and moral cleansing, Sachdeva *et al*. (2009) consider "*moral behavior as being embedded within a larger system that contains competing forces. Moral or immoral actions may emerge from an attempt to find balance among these forces*". The process is symmetric: every deviation from the normal behavior is subsequently balanced with either a more moral action (moral cleansing) or less moral action (moral licensing). In their experiment, Sachdeva *et al*. (2009) show that affirming a moral identity (participants were asked to write a self-relevant story containing positive traits) leads people to donate less to charities (moral licensing); when moral identity is threatened (story containing negative traits), generosity in donations to charity is a means to regain some lost self-worth (moral cleansing).

Our paper provides further evidence on this phenomenon of moral self-regulation in a dynamic context. We analyze data from an economic experiment where subjects play a sequence of 16 dictator games, each with a different randomly chosen recipient (anonymity conditions). All the games have the same structure and they are framed. Besides a blind (baseline) game, we use three types of frames regarding the information given about gender (male/female), income (poor/rich) and political preferences (right wing/left wing) of the dictator and the recipient, to generate 15 different environments. Each subject played the 16 games in a different random order to control for order effects.

This design tries to recreate the sequentially of decisions, to test the hypothesis of moral self-regulation that would lead individuals to reverse previous moral or immoral behavior. The alternative hypothesis is that subjects would always behave according to their moral standards and therefore we would observe no reversion.

Our estimation technique takes into account the dynamics of these actions; we estimate how a donation by each individual (d_{t-1}) affects the subsequent one (d_t). We find that donations over time follow an auto-regressive process of order one (AR(1)) with a negative coefficient.[2] We draw two important conclusions from this analysis:

i. the negative sign of the effect of the immediate past actions (d_{t-1}) on current choices (d_t) indicates that subjects reverse in every round what they did in the past;

ii. the length of the auto-regressive process (AR(1)) indicates that only the previous period affects present behavior. Hence, subjects tend to balance in period t what they did in period $t-1$.

Our result implies that self-regulation is not a long memory process, since only the previous period matters. This could be due to the fact that decisions in our experiment are not overly asymmetric so that only one period is sufficient to reverse what the subjects did in the past.

The rest of the paper is organized as follows. In Section II, we set the theoretical framework. Section III describes the experiment design and procedures, Sections IV and V contain the results and their robustness and in Section VI we present some concluding remarks.

II. THEORETICAL FRAMEWORK

This section presents a dynamic model for the paradox of moral self-regulation (Sachdeva, Iliev and Medin, 2009). In this theoretical framework, decisions with a moral content have to be taken over time and subjects self-regulate to achieve their optimal level of moral self-worth.

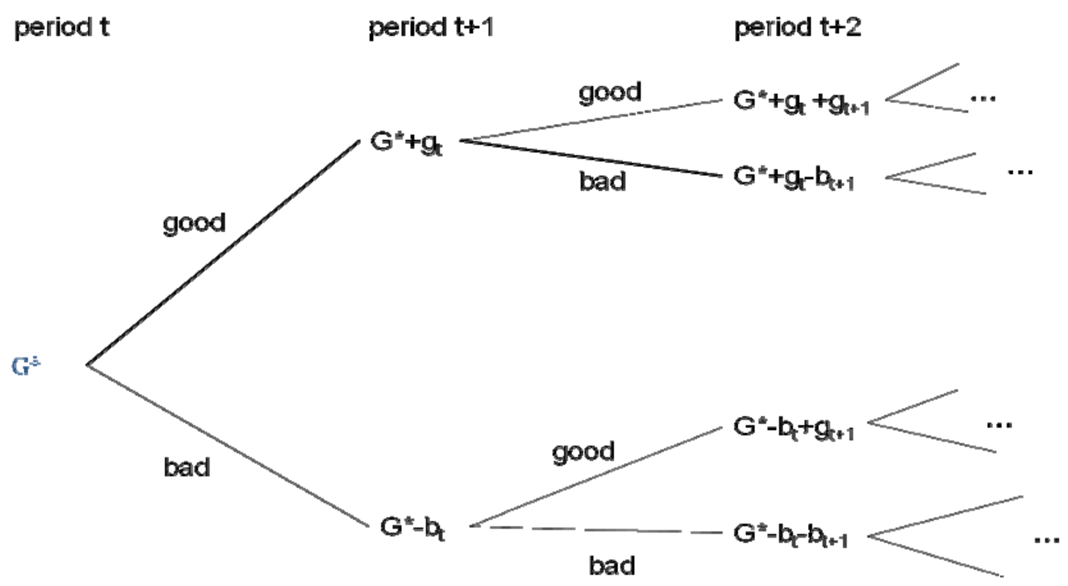
Assume that up to period t, a subject i is at her optimal level of goodness G^*, that is, she has taken decisions that have placed her in a situation where her moral self-worth is at the right level. This optimal level of goodness G^* is obtained taking into account the costs and benefits of moral self-worth, that is, the level of G^* maximizes $B(G)-C(G)$, the benefits minus the costs.

We assume that preferences concerning the level of goodness are single-peaked and symmetric around G* so that at each period t subjects minimize the distance $|G_t-G^*|$, where G_t is the moral self-worth at t.

In a dynamic context, this level G^* may be difficult to maintain since life requires difficult decisions with a moral content to be taken over time. To represent this, assume that at period t the individual must take a decision that will put her at a level of moral self-worth either higher than G^* or lower. We assume for simplicity that a single decision has to be taken each period, it cannot be avoided and that decisions are not neutral, that is, decisions always affect moral self-esteem.

We assume for simplicity that the decision at each period t is binary; the subject may either have good behavior, which increases goodness by $g_t>0$ or bad behavior which decreases it by $b_t>0$. Depending on the decision taken, she will enter period $t+1$ having a level of moral self-worth $G_t=G^*+ g_t$ or $G_t=G^*- b_t$.

Graph 1. Decision tree



As shown in Graph 1, the subject decides again in period $t+1$. If her decision was *good* in period t, she should choose *bad* in period $t+1$ as long as $G_{t+1} = G^* + g_t - b_{t+1}$ is closer to the optimal value G^* than $G_{t+1} = G^* + g_t + g_{t+1}$.

Note that if the decision *good* or *bad* is always symmetric, that is, if $g_t = b_t = g = b$ for all t, then the subject should always choose the decision opposite to the previous one, to get as close as possible to G^*.

Assuming that $g_t = g$ and $b_t = b$ for all t, what happens if the decisions are not symmetric (g≠b)? Take for example the case $b=3g$, that is, the cost of a bad action is three times the benefit of a good one. Then starting from G*, to minimize $|G_t - G^*|$ at each t, the subject's decisions should follow a regular pattern: (….gg b ggg b ggg b ……..), three good actions are always followed by a bad one.

More generally, if $b=ng$, *where* n is an integer and an even number, starting from G^* the optimal sequence of actions follows a regular pattern: *(n/2 actions g, one action b, n actions g, one action b, n actions g, one action b ……)*. If n is an odd number, the sequence is: *((n+1)/2 actions g, one action b, n actions g, one action b, n actions g, one action b ……)*. *If 1/n is an integer, and even number: (n/2 actions b, one action g, n actions b, one action g, n actions b, one action g ……)*. *If 1/n is odd: ((n+1)/2 actions b, one action g, n actions b, one action g, n actions b, one action g ……).*[3]

This result implies that individuals self-regulate to achieve their optimal level of moral self-worth *G** and this self-regulation follows a regular pattern. Whenever decisions with a moral content cannot be avoided, individuals will alternate bad and good actions over time.

In our experiment, we test whether these regular patterns predicted by the theory appear when subjects have to take sequential decisions involving moral self-worth.

III. The Experiment

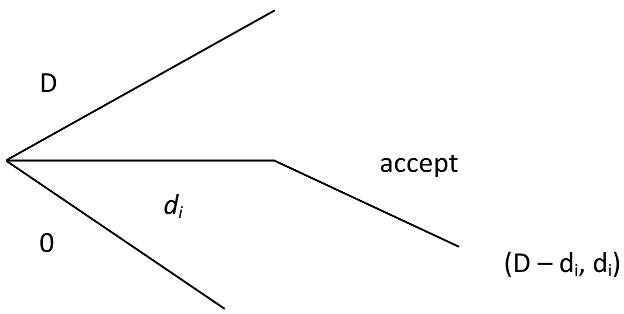
The dictator game

In the dictator game (Forsythe et al., 1994), the first player, "the proposer" (dictator), determines an allocation (split) of some endowment (such as a cash prize). The second player, "the responder" (recipient), simply receives the share of the endowment left by the proposer. The responder's role is entirely passive.

Formally, given an endowment of size D, the dictator must decide any value of $d_i \in [0, D]$ to pass to the recipient. Therefore the final distribution of benefits is a pair:

$$(D-d_i\ ,\ d_i)$$

where $D-d_i$ is the dictator's benefit. Since the Nash equilibrium is giving zero to the recipient, any strictly positive donation, $d_i>0$, is interpreted as pure altruism.



Participants

176 subjects distributed in four sessions participated in the experiment (dictators and recipients). We will focus only on the sample of 88 dictators (32 % of women) since recipients do not play any active role in our analysis. The participants were undergraduate students at the *Universidad de la República* (Uruguay). All of them were volunteers who answered a public call.

Procedures and materials

The subjects were given verbal and printed information: they had to take 16 decisions and each one was explained on one page of a printed booklet. They were not allowed to speak to one another and they were seated in such a way that they could not see the written responses of the other subjects.

The baseline treatment consisted of a standard dictator game in which each participant was a dictator or a recipient (the participants knew that no one would play both roles). The dictator had to allocate 10 bills of 20 Uruguay pesos (around 10 US dollars) between herself and a randomly chosen student placed in a different room. Following List (2007) instructions, the task was explained on one sheet of paper inside a printed booklet and the possible payoffs were presented on a line in which the subject had to mark her decision with a circle. The amount of money ranked from 0 pesos (left-end) to 200 pesos (right-end) and the donations were restricted to multiples of 20 including zero.

The rest of the treatments were identical to the baseline (blind) with the exception of the framing. In order to frame the task, we used information that participants gave at the moment they registered for the experiment: sex, income category and ideological category. This information was used to label the participants as women/men, rich/poor and right-wing/left-wing.[4]

In three treatments, the donor was told that the recipient would know the donor's sex, income category or ideological category, respectively. In six treatments, the donor knew one characteristic of the recipient (sex or income category or ideological category). In another six treatments, besides knowing one characteristic of the recipient, the donor was told that the recipient would know the game's framing (for example, the recipient would know that the donation was done from a woman to a man).

The entire booklet consisted of sixteen tasks that were presented in a different random order for each subject. This is an important characteristic of the design: as in each round the donors are facing different frames, even if all participants had the same preferences, we would not necessarily observe an equalizing pattern common to all subjects.

We paid only one decision (randomly chosen) to each dictator which avoids the effect of accumulation of earnings in the course of the session. Besides, the use of different recipients and frames at each decision helped to maintain subjects' interest. Notice that once a decision is taken, subsequent decisions by the same subject cannot actually hurt or help the same recipient. Thus, if the donor makes what he thinks is a selfish (generous) decision, the subsequent action will not compensate the prior recipient since the recipients are different individuals; any compensation effect affects exclusively moral self-worth with this design.

The money donated to recipients was delivered to them in a different session. Taking all the games into consideration, the average dictator's earnings in the 16 games were US$142.5 (7 bills) and, consequently, the average recipient's earnings were US$57.5 (3 bills).

IV. RESULTS

According to the theoretical framework described in Section II, we would expect a negative correlation between the donation at t and that at period $t+1$. We test this hypothesis in a dynamic panel data model where we estimate the donation at period t (d_t) as a function of past donation (d_{t-1}):

$$d_{it} = \alpha_i + \gamma d_{i,t-1} + x_{it}^{'}\beta + v_{it}, \quad i = 1,...,88 \text{ individuals}, \quad t = 1,...,16 \text{ rounds}$$

where α_i denotes the unobserved individual-specific time-invariant fixed effect[5]; x_{it} is the *it-th* observation of explanatory variables, in our case, treatment dummies and temporal trend; the disturbance terms v_{it} has zero mean, constant variance and is uncorrelated across time and individuals.

We use two-step GMM[6] estimators with the Windmeijer correction using lagged levels ($t-2$, $t-3$ and $t-4$) of the dependent variable as instruments (Arellano and Bond, 1991; Windmeijer, 2005).

Table 1 shows the results of three regressions. In the first one, the only covariate is the previous donation (d_{t-1}); in regression (2) we also include the treatment dummies and in regression (3) we add a temporal trend. In the three estimations, the coefficient of past donation (d_{t-1}) is negative, significant at 5% and less than one in absolute value. Besides, the trend is not significant. In the bottom part of Table 1 we show Arellano-Bond tests.[7]

Table 1: Moral cleansing and licensing

	(1)	(2)	(3)
Round (t)	-	-	0.195
			(0.430)
d_{t-1}	**-0.085**	**-0.088**	**-0.075**
	(0.035)	**(0.036)**	**(0.031)**
Constant	**61.125**	**48.081**	**45.157**
	(0.000)	**(0.000)**	**(0.000)**
Treatment controls	Not	Yes	Yes
Arellano-Bond serial correlation test	-0.635 (0.525)	-0.808 (0.419)	-0.695 (0.487)
Instruments	*40*	*43*	*44*
Sample Size	*1220*	*1220*	*1220*

p-values in parentheses.

The important result here is that donations follow a stationary AR(1) process with negative coefficient. Hence, subjects tend to balance a donation above the mean in a round with a donation below in the following round.

This result does not support the alternative hypothesis that subjects would always donate according to their moral standards and show consistent preferences for a given level of donation. On the contrary, the pattern of donations over time shows a self-regulation behavior and emerges as the result of a systematic process of dynamic equalization: moral licensing (being selfish after altruist) or cleansing (altruistic after selfish).

We also check if donations follow an AR(2) process. We find that the coefficient of d_{t-2} is not significant, whereas the coefficient of d_{t-1} is still negative and significant (show Table 2).

Table 2: Moral cleansing and licensing, with 2 lags

	(1)	(2)	(3)
Round (t)	-	-	-0.021 (0.938)
d_{t-1}	**-0.128 (0.073)**	**-0.135 (0.048)**	**-0.115 (0.045)**
	-0.055 (0.257)	-0.064 (0.165)	-0.052 (0.197)
Constant	**67.010 (0.000)**	**55.564 (0.000)**	**53.991 (0.000)**
Treatment controls	Not	Yes	Yes
Arellano-Bond serial correlation test	0.177 (0.860)	0.159 (0.873)	0.140 (0.888)
Instruments	*39*	*42*	*43*
Sample Size	*1130*	*1130*	*1130*

p-values in parentheses.

V. ROBUSTNESS

As a simple robustness test, we check whether our results change when we use different sample sizes. Table 3 shows the same regressions as before but using the last 12 periods (*t=5, 6, ... , 16*) and the last 8 periods (*t=9, 10, ... , 16*). Given that every individual played the 16 games in a different random order, we lose different treatments' observations for each individual.

Table 3: Robustness checks

	Rounds 5 to 16			**Rounds 9 to 16**		
	(4)	(5)	(6)	(7)	(8)	(9)
Round (t)	-	-	0.021 (0.931)	-	-	0.240 (0.625)
d_{t-1}	**-0.098 (0.061)**	**-0.098 (0.066)**	**-0.101 (0.036)**	**-0.119 (0.078)**	**-0.132 (0.042)**	**-0.137 (0.010)**
Constant	**63.335 (0.000)**	**51.237 (0.000)**	**50.744 (0.000)**	**65.612 (0.000)**	**50.090 (0.000)**	**47.245 (0.000)**
Treatment controls	Not	Yes	Yes	Not	Yes	Yes
Arellano-Bond serial correlation test	-0.592 (0.554)	-0.677 (0.498)	-0.724 (0.469)	-0.369 (0.712)	-0.536 (0.592)	-0.584 (0.559)
Instruments	37	40	41	25	28	29
Sample Size	1046	1046	1046	695	695	695

p-values in parentheses.

There are no remarkable differences when we compare results from Table 1 and Table 3. Hence, using all or only the final rounds of the experiment does not make any difference.

Lastly, Table 4 shows an additional robustness check. We estimate the AR(1) model -with controls- for a sample of 68 subjects randomly selected, that is, we drop 20 subjects. We repeat the exercise removing another 20 different subjects and finally we repeat the process a third time. Table 3 shows the estimated AR(1) coefficients for the three sub-samples (elimination #1, #2 and #3).

Table 4: Additional robustness checks

	AR(1) Coefficient	p-value	Sample Size
Removal of 20 participants			
elimination #1	-0.115	0.026	944
elimination #2	-0.105	0.057	941
elimination #3	-0.089	0.068	940
without "Blind"	-0.133	0.014	989
without "Constant"	-0.089	0.034	1094

Two additional robustness checks are shown at the bottom of Table 4. We estimate the AR(1) coefficients when observations from the baseline are not included; results are even stronger (*p-value=0.01*). We also run a model removing people who donate the same quantity in all rounds, and the results are identical to those obtained previously.

Our experimental results indicate that the coefficient of the participant's previous donation is significant and negative, which is consistent with our hypothesis that over time individuals self-regulate to attain the optimal level of self-worth.

VI. CONCLUSIONS

This research contributes to the literature that focuses on the role of moral cleansing and moral self-licensing on behavior. Our results show that donations do not have a trend over time in a dictator game setting. However this stability across time cannot be interpreted as the result of strong preferences for altruism. In contrast, this stability emerges as the result of equalization. In the estimations, the past donation (d_{t-1}) coefficient is always negative, significant and its absolute value is less than one-indicating that subjects who behaved nicely yesterday are selfish today and vice versa. In short, a systematic moral self-licensing and moral cleansing pattern emerges.

Our findings are related to the current theories of identity (Akerlof and Kranton, 2000). When decisions are not morally neutral, each decision affects the sense of identity and implies a deviation from the optimal level of moral self-worth, which requires a compensating subsequent decision. We have identified this self-regulation behavior empirically, which in our experiment takes the form of an autoregressive process of order 1 with a negative coefficient: a high (low) donation is followed by a low (high) donation. These results are consistent with moral licensing and moral cleansing.

References

Aguiar, F., Brañas-Garza, P., Espinosa, M. P. and L. Miller (2010). Personal identity. A theoretical and experimental analysis, *Journal of Economic Methodology* **17**(3): 261-275.

Akerlof, G. and R.E. Kranton (2000). Economics and Identity, *Quarterly Journal of Economics* **115**: 715–753.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations, *Review of Economic Studies* **58**: 277-297.

Battigalli, P. and M. Dufwenberg (2007). Guilt in games, *American Economic Review* **97**(2): 170-176.

Charness, G. and M. Dufwenberg (2006). Promises & Partnership, *Econometrica* **74**: 1579-1601.

Fehr, E. and U. Fischbauer (1999). The nature of human altruism, *Nature* **425**: 1579-1601.

Forsythe, R., Horowitz, J.L., Savin N.E. and M. Sefton (1994). Fairness in Simple Bargaining Experiments, *Games and Economic Behavior* **6**: 347-369.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games, *Journal of Political Economy* **115**(3): 482-93.

Merritt, A. C., Effron, D. and B. Monin (2010). Moral Self-Licensing: When Being Good Frees Us to Be Bad, *Social and Personality Psychology Compass* **4/5**: 344–357.

Sachdeva, S. Iliev, R. and D.L. Medin (2009). Sinning Saints and Saintly Sinners. The Paradox of Moral Self-Regulation, *Psychological Science* **20**: 523-528.

Windmeijer, F. (2005). A finite sample correction for the variance of linear efficient two-step GMM estimators, *Journal of Econometrics* **126**: 25-51.

Zhong, C. and K. Liljenquist (2006). Washing Away Your Sins: Threatened Morality and Physical Cleansing, *Science* **313**(5792): 1451-1452.

Appendix

The optimal sequence of actions

We assume that whenever the subject is indifferent between a good and a bad action, she chooses the one with the lowest payoff: b if $b<g$ and g if $g \leq b$.

Assume $b=ng$, n a positive integer and an even number. Starting from G^*, the subject has the choice between G^*-b or G^*+g, and she should choose G^*+g since it is closer to G^*. The same is true in the following periods up to period $n/2$. After $n/2$ periods, the subject is at $G^*+(n/2)g$. She is then indifferent between $G^*+(n/2)g$ and $G^*+(n/2)g-b=G^*+(n/2)g-ng= G^*-(n/2)g$, so that the next decision should be g since $G^*+(n/2)g-b$ is closer to G^* than $G^*+(n/2)g+g$.

Example. Assume G*=100, b=10 and g=2. Then n=5. The subject would follow the sequence (g,g,g,b,g,g,g,g,g,b,....): 100, 102, 104, 106, 96, 98,100,102, 104,...

When n is not an integer, the optimal sequence of actions takes a slightly more complicated form. For example, if $n=3.5$, the optimal sequence is (.....3 g’s, b, 4 g’s, b, 3 g’s, b, 4 g’s, b,…..).

The case $b=g/n$ follows by symmetry.