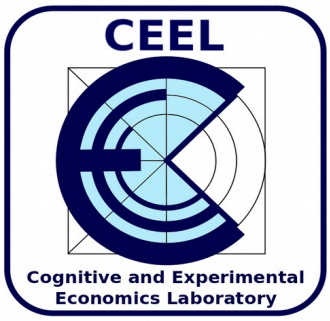


Sandro Casal, Francesco Guala and Luigi Mittone

On the Transparency of Nudges: An Experiment.

CEEL Working Paper 2-19

Cognitive and Experimental Economics Laboratory

Via Inama, 5 38100 Trento, Italy

http://www-ceel.economia.unitn.it
tel. +39.461.282313

On the Transparency of Nudges: An Experiment

Sandro Casal* Francesco Guala† Luigi Mittone‡

March 19, 2019

Abstract

We investigate the effects that transparency may have on people's reactions to a simple nudge. Using an incentivized task and eliminating possible confounds due to strategic reasoning, we test two behavioral predictions: (a) that increasing the quantity and quality of information affects significantly the efficacy of nudges; and (b) that people mind about being nudged and reverse their decisions when the behavioral policy is transparent. Our results indicate that transparency does not necessarily trigger reactance (people in general do not mind being nudged), but the quality and quantity of information can have a significant effect on the efficacy of a behavioral policy.

Keywords: Nudge, Overconfidence, Transparency, Reactance

1 Introduction

The proliferation of so-called nudges has generated a lively debate on the legitimacy and efficacy of behavioral policies. Although their precise definition is controversial, behavioral policies typically try to exploit people's cognitive limitations so as to induce decisions that are beneficial for themselves or their fellow citizens. While the governments of several countries have enthusiastically endorsed nudges, however, many scholars have found their normative foundations problematic. Most of the discussion so far has focused on *libertarian paternalism*, an approach defended by Richard Thaler and Cass Sunstein in various articles and a series of best-selling books.[1] Sunstein and Thaler defend two claims: (i) that behavioral policies help people make better decisions (i.e. they are 'paternalistic'), and (ii) that they do it without limiting people's freedom of choice (they are 'libertarian'). Although both claims have been disputed, in this paper we will focus mostly on the second one.

The critics of libertarian paternalism have pointed out that although nudges do not restrict the range of options that are available to decision-makers, they exploit cognitive biases that in practice may be difficult to overcome (e.g., Bovens (2009); Hausman and Welch (2010); Grüne-Yanoff (2012); Rebonato (2012)). Since people are usually unaware of their cognitive limitations, they are vulnerable to manipulation by governmental agencies and private firms. This lack of awareness is particularly problematic for libertarians, because it prevents people from exercising genuine freedom of choice. Although nudges may preserve *option*-freedom, they typically violate the *autonomy*-freedom of decision makers.

An obvious response to such a critique is that in many cases it is possible to increase the transparency of nudges. When people are informed about the existence of behavioral policies, they may autonomously decide whether to comply with the behavioral policy or not. Making nudges more transparent may help promote trust between citizens and policy-makers, while preserving the autonomy and responsibility of the former.

Such a solution however raises an important issue: if people do not like being manipulated by policy-makers, they may "rebel" against nudges, once they have become aware of their existence (a phenomenon known as reactance in the psychological literature. See Brehm and Brehm (2013) for a comprehensive discussion). In such cases, the policy-maker would face a dilemma: she could either preserve the efficacy of the behavioral policy, at the expense of people's autonomy; or she could respect people's autonomy, but face the risk of making the policy ineffective or counterproductive.

Notice that the threat of reactance is entirely an empirical issue. Perhaps people do not mind being nudged: in such a case, their indifference could be interpreted as a mandate to the (benevolent) policy-maker to steer behavior in the right direction. But perhaps they do mind, and implementing the nudge would constitute a serious breach of citizens' trust in the transparency of government intervention.

This paper contributes to the debate on behavioral policy and libertarian paternalism by investigating the reactions of experimental subjects when they are provided with different levels of information about a behavioral policy. In particular,

*Department of Economics and Management and Cognitive and Experimental Economics Laboratory, University of Trento (Italy).
Email: sandro.casal@unitn.it
†Department of Philosophy and PhiLab, University of Milan (Italy).
Email: francesco.guala@unimi.it
‡Department of Economics and Management and Cognitive and Experimental Economics Laboratory, University of Trento (Italy).
Visiting professor at Lappeenranta University of Technology (Finland).
Email: luigi.mittone@unitn.it

[1] See e.g., Thaler and Sunstein (2003, 2009), Sunstein (2006, 2013, 2014)

the experiment is meant to test two predictions: (a) that increasing the quantity and quality of information affects significantly the efficacy of nudges; (b) that people care about being nudged, and reverse their decisions when the behavioral policy has been made transparent.

Although this is not the first paper to address these issues, the evidence so far has been mixed and has failed to provide clear answers. On the one hand, data from properly controlled and incentivized experiments are scarce. On the other, transparency and reactance have been studied in situations in which their effects may be confounded with strategic reasoning. As far as we know, this is the first attempt to collect experimental data with an incentivized task, in an environment where the effect of the quality and quantity of information on individual decisions is not confounded by strategic considerations.

The paper is organized as follows: in the next section we summarize and discuss the existing literature on transparency and reactance. We then proceed to illustrate the main features of our experiment (section 3) and to analyze the experimental data (section 4). Section 5 concludes with a discussion of the main results.

2 Background

Experimental data on reactance to behavioral policies is scarce, mostly unpublished, and the evidence is mixed. Moreover, much of the behavioral evidence comes from non-incentivized tasks.

Arad and Rubinstein (2018) report the results of a non-incentivized survey conducted in three countries (Germany, Israel, and the US). People were asked hypothetical questions about their attitude about and willingness to participate in a governmental program aimed at increasing saving. The survey tried to figure out if transparency could induce some subjects to reverse their decision – for example, to opt out of a program they had been previously nudged to opt in. Arad and Rubinstein report that a significant proportion of individuals (up to two-thirds, in one treatment) expressed a negative attitude toward the nudge, with large variations across countries and types of nudge. They also find that awareness makes a significant proportion of people opt out of the program (up to 30%, in one treatment), again with significant cross-country variations.

Loewenstein et al. (2015) have studied hypothetical decisions about end-of-life treatment directives. The nudge takes the form of a default option that is implemented automatically unless people decide to opt-out. The treatment is the transparency of the default, and the explicit offer to reverse a decision previously made without transparency. The task is hypothetical and non-incentivized. Lowenstein and colleagues find that the default does not have an effect on decision (which they interpret as evidence that 'the respondents knew well their overall goals for care'). They also find that 15-20% of subjects change their initial directive after they have been made aware of the nudge (which they interpret as favorable to the hypothesis that information does not have a significant effect). Given the hypothetical nature of the task, and the lack of efficacy of the nudge, however, it is difficult to derive any conclusions from this experiment.

Jachimowicz et al. (2016) have asked subjects about the use of environmentally-friendly materials in a hypothetical house renovation project. Their nudge shifts participants toward the choice of more environmentally-friendly materials. Information about the existence of the nudge affects the behavior of 40% of participants (when the nudge is not clearly counter-preferential, however, the proportion declines to 19%). These data seem to suggest that transparency may have a significant impact of behavior, although once again the results are not fully reliable due to lack of incentives.

Petrescu et al. (2016) report a survey conducted in the UK and USA, aimed at studying the acceptability of government interventions to reduce the consumption of sugar-sweetened beverages. The study probes people's reactions when the intervention is described as affecting behavior via conscious processes or non-conscious processes. While respondents in general prefer education to nudges, and nudges to taxation, they do not consider nudges less acceptable when the intervention modifies behavior unconsciously.

Sunstein (2016) provides a wide survey and critical discussion of people's attitudes toward nudging. He interprets the literature as demonstrating that attitudes vary greatly depending on political orientation and people's perception of the intentions of the policy-maker. Nudges that support deliberative thinking and conscious decision-making seem to be more appreciated than those that exploit subliminal or subconscious mechanisms. On transparency, in particular, Sunstein claims that it does not matter much – but his claim is backed only by the study of Loewenstein et al. (2015), which, as we have seen, is not particularly convincing. Finally, Sunstein mentions reactance as a topic that deserves to be further researched in the future.

The most recent published paper on this topic, by Bruns et al. (2018), studies contributions to a global public good in an experimental setting incentivized with real monetary stakes. The experiment focuses on two factors: transparency about the behavioral *influence* vs. transparency about the *purpose* of the nudge. Although they report a significant effect of the nudge, Bruns and co-authors find that transparency (of either kind) has no significant effect on contribution rates. Their design however raises a major worry, namely that the effect of information about the nudge may lead people to change their decisions *for strategic reasons* (because they anticipate that the information may affect the behavior of others). It is well known, in fact, that people's decisions in a public goods game may depend on social preferences and on mutual expectations of contribution (e.g.Chaudhuri

(2011)). Individuals who are willing to reciprocate the cooperation of others, in particular, may interpret the task as a coordination game. A nudge in such circumstances may be perceived as a signal, which the subjects follow if they believe that others have similar preferences, in order to facilitate coordination. What Bruns and co-authors interpret as increased transparency, therefore, may be merely a case of salient signaling.

3 Experimental design

To avoid complications with social preferences and strategic considerations, we study reactance behavior using an individual decision task under uncertainty. And in order to elicit more reliable data, we incentivize the task using real money. The experiment implements three important features of behavioural policy cases.

1. The decision-maker pursues an observable and quantifiable individual goal
2. She is faced with a conflict between an appealing (but worse) alternative and a less attractive (but better) option.[2]
3. The choice architecture may be manipulated to channel behaviour towards one of the alternatives.

In our experiment decision makers are exposed to one of the most robust biases documented in the psychology and behavioral economics literature: the so-called *optimistic bias* (Weinstein and Klein, 1996; Chapin and Coleman, 2009). The optimistic bias is a tendency to make predictions about one's own future well-being that are more positive than it would be rational to make. This tendency is particularly strong when the decision makers think that their future well-being depends on their actions. Once overconfidence is experimentally triggered, it should be possible to introduce a nudge aimed at reducing its negative effects.

To achieve full control on individual preferences, we engage experimental subjects with an abstract task that does not resemble closely any concrete real-life problem they may be familiar with. In a sense, thus, our design is a compromise between the real-world situations described by the literature on nudging and the constraints imposed by the artificial environment of the laboratory. This approach allows to measure with precision the effectiveness of the nudge and to control, in a second stage, the effect of transparency on behaviour.

In our experiment each subject must try to predict her performance in a simple but unfamiliar puzzle game. The final payoffs depend on the accuracy of each subject's prediction, with a maximum penalty for those who overestimate their performance (they do worse than they have predicted). Our goal is not so much to measure performance, however, but to study the effects of various nudges on subjects' predictions. Each nudge is indeed accompanied by a different explanation or justification. More precisely, we compare behavior in a Baseline condition in which predictions are made freely, against four treatments: a 'Default' condition with a non-transparent nudge, a 'Warned' condition with a nudge accompanied by cautioning, a 'Justification' condition with a nudge backed by a reason to comply, and a 'Transparent' condition with a fully explained nudge. In the next subsections we describe the task and the treatments in more detail.

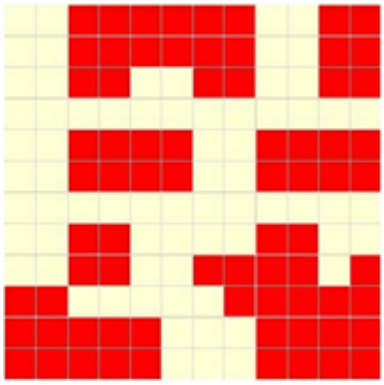


FIGURE 1: Example of a Target

3.1 The grid task

Subjects were shown a target figure ('Target') on screen, constituted by a 12x12 grid. Each square or pixel was either colored (red) or blank (see Figure 1 for an example).

Next to the target, the subjects would see an empty grid (called 'Build'), and sixteen 'blocks' constituted by 2x2 pixels with all possible combinations of red or beige pixels (see Figure 2).

FIGURE 2: Examples of 'blocks'

The goal of this task (henceforth called the 'grid task') was to replicate the Target using the blocks, which could be selected and placed (or removed) on the 'Build' grid, sequentially and repeatedly. Individual payoffs depended on the number of Targets the subject was able to replicate successfully,[3] within a given time limit. The game would end or continue after each round, depending on the subject's

[2] This is meant to reproduce the main characteristic of a wide number of nudges discussed in the literature. A classic example of "undesirable" behavior that may be corrected through the implementation of a nudge is the consumption of trash food. In that case, the conflict is between the short-period goal of enjoying cheap and tasty food and the long-term goal of preserving good health.

[3] A round was 'successful' whenever the Target was replicated in the Build grid within a margin of error of 5 pixels.

performance. The available time decreased as the rounds progressed, starting from a maximum of three minutes (to replicate the Target in the first round) to a minimum of one minute and fifteen seconds (in the eighth round). When a subject failed to replicate the Target within the time limit, the game ended. When the Target was replicated, the game proceeded to the next round (for a maximum of eight rounds).

All this information was available in the instructions provided at the beginning of the experiment (see the Appendix). After reading the instructions, each subject was asked to predict the number of consecutive Targets (or rounds) that she would be able to complete successfully (the 'Objective'). This forecast would later determine her base payoff: if the subject was unable to achieve the Objective, she would earn nothing. If the subject achieved the Objective, she would earn 2.00 euro for each Target she had forecasted and replicated. Every Target successfully replicated *after* attaining the Objective would earn the subject an extra sum (50 cents). To facilitate comprehension, a table of monetary payoffs was provided with the instructions.

Monetary payoffs were designed not only to incentivize effort (the earnings increased with the number of replicated Targets), but also (indeed, especially) to incentivize accurate forecasts. However, given subjects' limited experience with the task, we expected many forecasts to be inaccurate. As we shall see, this was confirmed by the data collected in the experiment. Subjects' inaccuracy (or sub-optimal behavior) thus gave us an opportunity to study the effects of different nudges in different conditions, each one explicitly designed to probe subjects' reactions to the transparency of the manipulation.

Baseline condition

Subjects played the grid task as explained above. In order to make their forecasts, they had to enter a number (from 1 to 8) in an empty box. The on-screen text simply said: "Please declare your Objective".

Figure 3 summarizes the forecasts (on the X-axis) made by subjects in the Baseline sessions, as well as their actual results (the number of replicated figures, on the Y-axis). Intuitively, the bubbles above or on the dashed diagonal line represent 'successful' subjects (i.e., subjects who have replicated at least their respective forecasted number of figures), while those below the line represent failed attempts to attain the forecasted number of figures. Errors in participants' predictions were prevalently due to overestimation of their performance, which confirms the presence of an optimistic bias. Looking at Figure 3 we can see that 27 participants in the baseline treatment overestimated their actual performance, while 17 participants correctly forecasted or underestimated their performance.

We used these data to identify the manipulation (nudge) that we would use in the other experimental sessions. Our

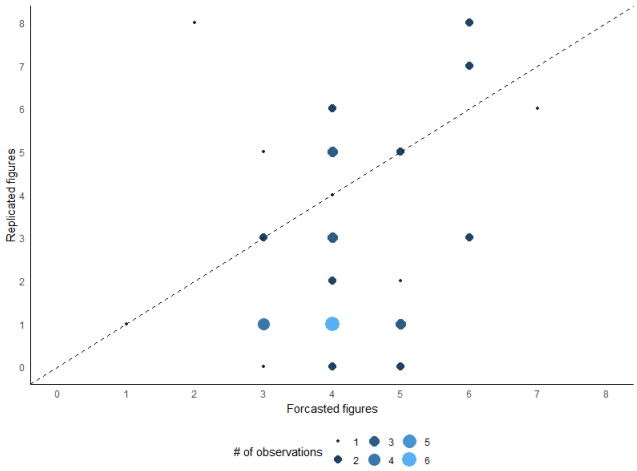


FIGURE 3: Number of observations for each potential combination of forecasted figures (Objective) and replicated figures in Baseline.

basic nudge consisted in proposing a default forecast that the subjects could change, if they wanted to. In light of the data collected in the Baseline condition, we chose a default that (i) could significantly change the behavior of subjects, leaving some room for maneuver; (ii) could affect the highest number of subjects; (iv) would be easy to understand, and (v) would be easy to justify or explain to the subjects. Given that the average forecast in the baseline was 4.23, we chose two Targets as our default forecast. Such a forecast was easy to justify, since those subjects who forecasted two Targets or less were always successful in the Baseline condition; it could affect a large number of participants, since only a few predicted less than three Targets; and left enough room for maneuver to observe significant behavioural variations.

Default condition

This condition differed from the Baseline in one respect only: the subjects had to make their forecast by selecting a number (from one to eight) from a drop-down menu which featured the pre-selected option (forecast) of two Targets. The subjects could either accept the default by clicking "continue", or use the drop-down menu to change their forecast. If they did so, a pop-up window would appear, asking to confirm their choice.

In all the remaining treatments, subjects selected their forecast exactly as in the Default condition, except that the preselection was combined with different explanations.

Justified condition

The preselection was explained as follows:

The Objective has been preselected at 2 because in previous experimental sessions those subjects who declared less than

3 targets had a 100% success rate, whereas the rate of success declined to less than 36% among those who declared 3 or more targets as their Objective.

The statement described truthfully the data obtained in our Baseline condition, and was intended (i) to make the subjects aware of the nudge, and (ii) to explain the rationale of the chosen default forecast.

Warned default condition

In this condition the explanation was formulated as follows:

The Objective has been preselected at 2 because, not knowing the individual skill of each participant in replicating the targets, the preselection at 2 is a precautionary choice. However, it's important to know that the preselected Objective at 2 may not reflect your real skill and therefore could make you lose the opportunity to earn more money, in case you were capable of replicating more targets.

This statement was meant to highlight the potential negative 'side effects' of the default.

Transparent default condition

This condition combined the information provided in the Justified condition and in the Warned condition, with the aim of making the nudge as transparent as possible:

The Objective has been preselected at 2 because in previous experimental sessions those subjects who declared less than 3 targets have had a 100% success rate, whereas the rate of success declines to less than 36% among those who have declared 3 or more targets as their Objective. Since we do not know the individual skill of each participant in replicating the targets, the preselection at 2 is a precautionary choice that, on the basis of statistical data, should guarantee that everyone will attain (and perhaps do better than) the forecasted Objective, avoiding the risk of earning nothing. However, it's important to know that, being based on statistical data, the preselected Objective at 2 may not reflect your real skill and therefore could make you lose the opportunity to earn more money, in case you were capable of replicating more targets.

The main goal of this condition was to offer a comprehensive account of the reasons as well as the risks involved in choosing the preselected default. Subjects would thus possess all the relevant information and could deliberate autonomously whether to follow the nudge or not.

Overall, these five conditions allowed to test some of the behavioral predictions outlined in the introductory section. In particular, we wanted to check whether the transparency of a nudge manipulation had a (positive or negative) effect on the efficacy of the nudge itself, and whether it could trigger reactance. The design fulfills the three conditions (a, b, c) highlighted at the beginning of this section. The abstract puzzle task provides a measurable objective function for individual decision-makers. The forecasts may be 'good' or 'bad' depending on subjects' vulnerability to the overconfidence bias. Finally, the choice architecture may be designed in such a way as to nudge subjects' decisions towards better (i.e. more accurate) or worse (overconfident) forecasts.

3.2 Participants and Procedures

The data were collected at the Cognitive and Experimental Economics Laboratory (CEEL) of the University of Trento, during ten sessions run in five separate occasions. Each session lasted approximately 45 minutes and the tasks were implemented using o-Tree (Chen et al., 2016). Overall, 203 subjects participated in the experiment, recruited from the student population of the University of Trento. Male subjects constituted 45.32% of the sample. The average earning for the whole experiment was 8.50 euro (including a show-up fee of 3 euro and the payoff of the 'Bomb Risk Elicitation Task' (BRET) explained below).

Subjects entered the lab and were seated randomly at their computer terminals, separated by partitions. The experimental instructions were read aloud by an assistant, while participants followed the text on their screens. After all clarification questions were answered, the experiment began. Before the main task, we elicited subjects' risk attitudes using the so-called 'Bomb Risk Elicitation Task' (BRET), a measurement tool devised by Crosetto and Filippin (2013) that is becoming increasingly popular in experimental economics.[4] In this task, subjects are presented with a 10x10 square in which each cell represents a box: 99 boxes contain 0,03 euro, while one contains a bomb. Each subject chooses how many boxes to collect ($k_i \in \{1, 100\}$) knowing that if the bomb is collected the earnings will be zero: the position of the bomb ($b_i \in \{1, 100\}$) is randomly determined after the subject's choice. If $k_i \geq b_i$, it means that the subject collected the bomb, which by exploding wipes out the earnings. In contrast, if $k_i < b_i$ the subject receives 0,03 euro for every box collected. The chosen number of boxes provides a measure of risk attitude: the lower the number, the more risk averse the subject. $k_i = 50$ represents a risk neutral choice. The results of the BRET task were not announced immediately, so as to avoid any endowment effect.

4 Results

Table 1 includes descriptive statistics of the two main variables of interest, namely, the distribution of forecasted Targets (or Objectives) and the distribution of replicated Targets.

[4] Among other advantages, the BRET is highly intuitive (it requires only minimal numeracy), and is unaffected by loss aversion or violations of the reduction axiom.

TABLE 1: Distribution of forecasted and replicated figures, for each condition.

	Number of figures								
	0	1	2	3	4	5	6	7	8
					Forcasted				
Baseline	-	1	1	8	19	8	6	1	0
Default	-	0	8	10	10	5	3	2	0
Justified	-	0	12	9	8	5	5	1	0
Warned	-	0	7	2	13	8	7	1	0
Transparent	-	0	13	7	12	8	3	0	0
					Replicated				
Baseline	5	14	3	7	1	6	3	2	3
Default	6	10	1	4	8	2	3	4	0
Justified	5	5	3	7	3	5	2	7	3
Warned	2	13	3	3	3	4	3	3	4
Transparent	4	7	4	6	8	6	2	4	2

Notice that subjects' performance in the grid task (the number of replicated figures) in principle should not be affected by the experimental treatments. In contrast, if the nudge is successful, we should expect both the forecasts and, subordinately, the payoffs to vary across the conditions. It is worth underlining that we are mainly interested in the effects of the treatments on the number of forecasted target and, only as a consequence, in the degree of maximization of the payoff function assigned to our participants. An analysis of the actual average payoff is not crucial, since the degree of success in terms of monetary payoffs depends on the distribution of skills in the population: the more dispersed are the latter, the less efficient should the nudge be, in terms of average monetary payoff maximization.[5]

We begin to analyze the results focusing on the number of replicated Targets, and then look at the main variable of interest, namely, the number of Targets forecasted in the various conditions.

[5]The gap between outcomes and nudged behaviours reflects one of the least discussed characteristics of nudging. Imagine a young person who loves to smoke cigarettes more than anything else, but is induced to stop smoking by a behavioural policy. On average, people of this kind will lose years of pleasure from smoking but will get a longer and healthier life (i.e., the nudge in a sense is successful). Imagine, however, that this specific person has a genetic make-up that protects it completely from the negative consequences of smoking. In this case the net balance between loss of pleasure and health benefits would be minimal or even negative. The overall efficiency of the behavioural policy will then depend on the distribution of genes in the population. In reality a genetic "shield" of this kind is probably very rare, so the introduction of a nudge against smoke is likely to produce an average positive effect. Nevertheless, this may not always be the case and, on a purely individual basis, this is certainly not always the case. The point of this example is that in our experiment we are satisfied if we obtain on average an improvement in monetary payoffs, even if the difference is small.

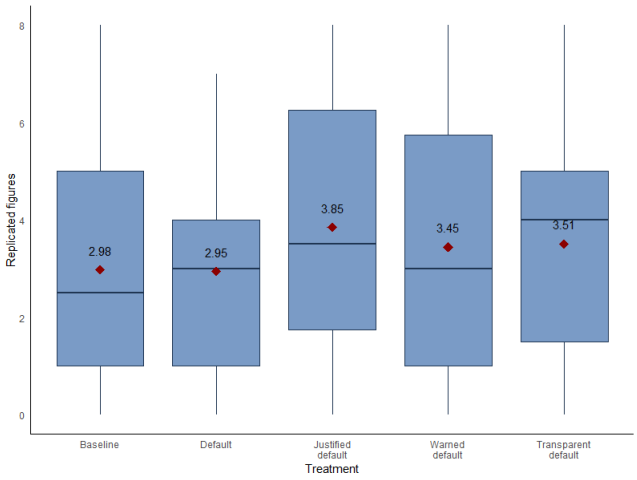


FIGURE 4: Distribution of replicated Targets across the experimental conditions. Diamond point represents the average value (also reported numerically in the box-plot).

Figure 4 represents the average (diamond dots, also numerically reported within the respective plot), the median, and the variance in the number of replicated Targets for each condition. While subjects seem to have performed slightly better in the Justified and Transparent conditions, this effect is probably due to mere chance. As expected, a series of Mann-Whitney tests does not detect any significant differences between the various conditions (for all binary comparisons, p-values > 0.1281).

Figure 5 represents the number of forecasted Targets (Objectives). While the average and median values seem rather similar, the distributions are sufficiently different across some conditions to pass conventional significance levels in a Mann-Whitney test. The data in the Default, Justified, and Transparent conditions, in particular, are significantly different from those in the Baseline (p-values, respectively, of 0.059, 0.025, and 0.020) but a significant difference is not detected when comparing Baseline and Warned conditions (p-value of 0.794). This indicates that the nudge was always successful except when the default was accompanied by a only warning about its potential negative effects. In the latter case, subjects did not declare a significantly lower number of figures than in the Baseline. This in turn suggests that the *quality* and *quantity* of information affects subjects' attitude toward the nudge. Notice, finally, that the additional information provided in the Justified condition did not seem to have a substantial effect compared to the bare Default (p-value = 0.619). Similarly, forecasts in the Transparent condition did not differ significantly from those in the Default (p-value = 0.627) and Justified conditions (p-value = 0.966).

The first important implication of these data is that there is no evidence of reactance effects. Making subjects aware of the existence of a Default does not make them change their

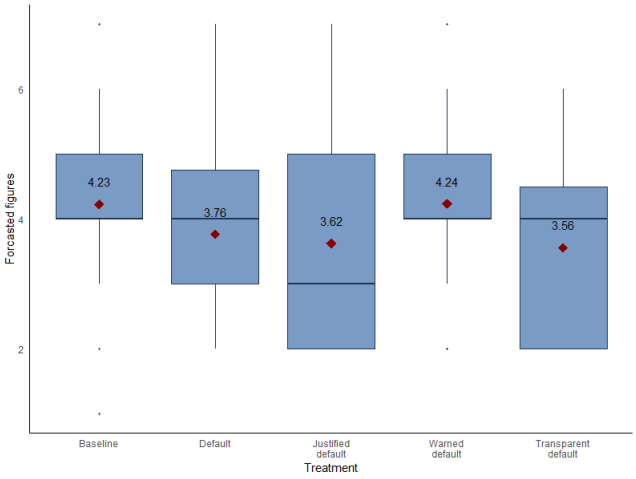


FIGURE 5: Distribution of forcasted Targets across the experimental conditions. Diamond point represents the average value (also reported numerically in the box-plot).

behavior. The quality and quantity of information, however, do matter. Warning the subjects about the possible negative implications of the Default seems to neutralize the effect of the nudge. Providing a comprehensive account of the possible implications of the nudge, instead, does not deter subjects from following the default. In general *transparency matters*, and how you make the nudge transparent matters a lot.

We report now the results of a series of regressions using the forecasted number of Targets as our dependent variable. In Model (1) of Table 2 we use four dummy variables that identify the respective conditions (Baseline is the omitted variable). Notice that only the Justified and Transparent conditions pass the conventional significance levels. The Default condition moves in the expected direction – i.e., it reduces subjects' number of forecasted figures, but only approaches conventional significant levels (p-value = 0.12), while the condition with Warning does not have a significant impact on forecasts.

Since transparency was attained by communicating the possible risks of the nudge, it is natural to ask whether the effects of transparency interact with individual attitudes toward risk. To answer, we use a simple parameter elicited using the BRET, namely the number of 'boxes collected' by each subject. The regression reported in Model (2) of Table 2 indicates that individual propensities toward risk influence the forecasts made by experimental subjects in all conditions. Intuitively, the less risk-averse the subject is, the more Targets she forecasts.

Combining the previous two regressions we obtain a 'full model' with treatment and risk-attitude as explanatory variables. The results in Model (3) of Table 2 indicate that the inclusion of risk attitudes makes the Default condition pass the 10% significance level. All the other treatment variables remain significant, except the default condition with Warning.

TABLE 2: Main effects on the forecasted number of figures.

	Dependent variable:		
	Objective		
	(1)	(2)	(3)
Warned	0.010		0.002
	(0.300)		(0.296)
Justified	−0.602**		−0.668**
	(0.296)		(0.293)
Transparent	−0.669**		−0.740**
	(0.291)		(0.288)
Default	−0.464		−0.496*
	(0.300)		(0.296)
BRET		0.011**	0.012***
		(0.005)	(0.005)
Constant	4.227***	3.413***	3.711***
	(0.204)	(0.233)	(0.281)
Observations	203	203	203
R^2	0.046	0.024	0.079
Adjusted R^2	0.027	0.019	0.055
Signif. codes:	*p<0.1; **p<0.05; ***p<0.01		

Notes: Linear regression model with std. errors in parenthesis. The dependent variable ('Objective') is the forecasted number of figures by the subjects. Baseline condition is the omitted category. BRET is the number of boxes collected in the BRET.

Finally, we investigate the effect of two observable characteristics, namely gender and faculty affiliation. Our entire sample includes 111 female subjects (gender=1 in Table 3) e 92 male subjects (gender=0). The distribution is slightly unbalanced in the Transparent and Default conditions (approximately 63% women, in both). Many subjects are students of business and economics, but there is also a substantial minority of law students. Combining all non-economics students in a single category, they account for 53% of the

sample. The distribution is unbalanced in favor of business and economics in the Warning condition (76%), and in favor of other faculties in the Justified and Transparent conditions (72% and 67% respectively).

Table 3 reports the impact of gender and faculty affiliation on the number of forecasted Targets.[6] The effect of gender is strong and negative: being a female reduces the number of forecasted Targets significantly. Studying economics, in contrast, does not seem to matter: the effect of this variable is not even close to significance according to standard statistical criteria.

TABLE 3: The effects of observable variables on the forecasted number of figures.

	Dependent variable:
	Objective
gender	-0.639^{***}
	(0.189)
econ	0.168
	(0.189)
Constant	4.152^{***}
	(0.171)
Observations	203
R^2	0.060
Adjusted R^2	0.051
Signif. codes:	*p<0.1; **p<0.05; ***p<0.01

Notes: Linear regression model with std. errors in parenthesis. The dependent variable ('Objective') is the forecasted number of figures by the subjects. Gender is a dummy equal to 1 when the subject is a female (0 otherwise). Econ is a dummy equal to 1 when the subject is an economics student (0 otherwise).

We conclude our analysis, for completeness, discussing the distribution of payoffs (only related to the grid task) in the five experimental conditions. Figure 6 summarizes the average and median earnings, as well as the variance for each condition. Running a series Mann-Whitney test (Table 4) we find no significant differences, except in the Justified condition (compared to the Baseline), and two borderline cases: Transparent vs. Baseline, and Justified vs. Default.

[6] Although we elicited individual faculty affiliations, as already introduced the large majority of the subjects attended a degree in business and economics, hence we classify subjects as either an economics student or not. Thus, *econ* takes value equal to 1 for students of business and economics and 0 otherwise.

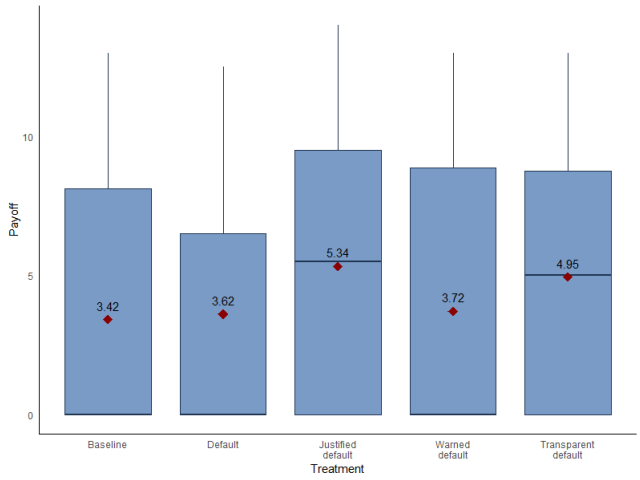


FIGURE 6: . Distribution of payoffs across the experimental conditions. Diamond point represents the average value (also reported numerically in the box-plot).

TABLE 4: Non-parametric tests (p-values of Mann-Whitney tests) on differences in payoffs across conditions.

	Default	Justified	Warned	Transparent
Baseline	0.864	0.060	0.808	0.102
Default	-	0.109	1	0.194
Justified	-	-	0.125	0.730
Warned	-	-	-	0.200

As anticipated, the data on final payoffs are difficult to interpret. Subjects' performance in the Justified and Transparent conditions may be explained by the fact that the nudge lowered the Objective, facilitating success; but it may also be due to the fact that some subjects invested more effort in completing the task. The problem of course is that those who were not influenced by the nudge may have been able to attain their Objective in some conditions, but not in others. However, we cannot identify with certainty at the individual level those subjects who were influenced by the nudge, and thus we cannot control for this effect with much confidence. As a mere hypothesis, we can guess that those who did not change the default value were nudged (although some of these would have probably forecasted two Targets anyway) and see how many of these were successful in the task. These data are reported in Table 5.

Taking them with a big pinch of salt, we notice that 'non-nudged' subjects (those who did not forecast two Targets) were more successful in the Justified and Transparent conditions. This may also explain why payoffs are higher on average in these two treatments.

TABLE 5: Non-parametric tests (p-values of Mann-Whitney tests) on differences in payoffs across conditions.

	Default	Justified	Warned	Transparent
		nudged		
# obs.	8	12	7	13
% of success	25	50	42.85	53.84
		non-nudged		
# obs.	30	28	31	30
% of success	50	71.42	38.71	66.67

5 Summary and conclusions

One of the most powerful objections leveled against behavioral policies concerns the violation of citizens' right to make autonomous decisions in important matters such as saving or healthcare. This objection however rests on two important empirical assumptions: first, it presupposes that people care about being nudged; and second, that they would change their behavior significantly if they knew that they are being nudged. If these two assumptions do not hold, then the objection loses most of its bite: behavioral policies would be justifiable in those contexts in which nudges are made transparent and people do not mind about being steered by policy-makers. The experimental literature is just beginning to investigate the effect of transparency, but the existing studies have failed to come up with convincing evidence. While most data so far have been elicited in experiments without monetary payoffs, the only incentivized study makes use of a task (a public good game) that is unable to separate strategic behavior from the effect of the transparency of nudge.

The experiment reported in this paper overcomes these limitations, and investigates the effect of transparency across four conditions that vary according to the quality and quantity of information. Subjects are nudged towards making cautious forecasts in a simple puzzle game, and are given different types of information about the rationale behind the nudge. This information inevitably may be read as a *reason* to comply (or not) with the nudge, and must be treated carefully. The data suggest that transparency matters – it can change behavior – depending on the type of information (reason) that is provided. When subjects are only warned about the risks that the nudge may entail, we observe a small reactance effect. But when the nudge is fully explained and justified, the reactance effect seems to disappear.

These results may seem, in a way, unsurprising: they suggest that people pay attention to the advice they receive, and are sensitive to the reasons for or against a particular course of action. If this is the case, however, one may wonder whether the very idea of nudging (using subliminal manipulations of behavior) should be set aside. Exposing people to all the pros and cons (risks and opportunities) may be an equally effective way to stimulate sensible decisions, without violating the autonomy of decision-making. Of course we do not claim that we have conclusively resolved this issue by means of a single experiment: the efficacy of nudges and the effects of transparency may vary across contexts and decision tasks, so we will need more data collected in different environments in order to draw any firm conclusion. Our main goal here is merely to show that such questions can be tackled by means of appropriately controlled experimental tasks, using a variety of strategies to increase the awareness of subjects with respect to the manipulation that is being used. What matters is not *that* the nudge is transparent, but *how*.

References

Arad, A. and Rubinstein, A. (2018). The people's perspective on libertarian-paternalistic policies. *The Journal of Law and Economics*, 61(2):311–333.

Bovens, L. (2009). The ethics of nudge. In *Preference change*, pages 207–219. Springer.

Brehm, S. S. and Brehm, J. W. (2013). *Psychological reactance: A theory of freedom and control*. Academic Press.

Bruns, H., Kantorowicz-Reznichenko, E., Klement, K., Jonsson, M. L., and Rahali, B. (2018). Can nudges be transparent and yet effective? *Journal of Economic Psychology*, 65:41–59.

Chapin, J. and Coleman, G. (2009). Optimistic bias: What you think, what you know, or whom you know? *North American Journal of Psychology*, 11(1):121.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14(1):47–83.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Crosetto, P. and Filippin, A. (2013). The "bomb" risk elicitation task. *Journal of Risk and Uncertainty*, 47(1):31–65.

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. *Social Choice and Welfare*, 38(4):635–645.

Hausman, D. M. and Welch, B. (2010). Debate: To nudge or not to nudge. *Journal of Political Philosophy*, 18(1):123–136.

Jachimowicz, J. M., Duncan, S., and Weber, E. U. (2016). Default-rejection: The hidden cost of defaults. Unpublished.

Loewenstein, G., Bryce, C., Hagmann, D., and Rajpal, S. (2015). Warning: You are about to be nudged. *Behavioral Science & Policy*, 1(1):35–42.

Petrescu, D. C., Hollands, G. J., Couturier, D.-L., Ng, Y.-L., and Marteau, T. M. (2016). Public acceptability in the uk and usa of nudging to reduce obesity: the example of reducing sugar-sweetened beverages consumption. *PLoS One*, 11(6):e0155995.

Rebonato, R. (2012). *Taking liberties: A critical examination of libertarian paternalism*. Palgrave Macmillan.

Sunstein, C. (2006). Preferences, paternalism, and liberty. *Royal Institute of Philosophy Supplements*, 59:233–264.

Sunstein, C. R. (2013). *Simpler: The future of government*. Simon and Schuster.

Sunstein, C. R. (2014). *Why nudge?: The politics of libertarian paternalism*. Yale University Press.

Sunstein, C. R. (2016). Do people like nudges? *Administrative Law Review, Forthcoming*.

Thaler, R. H. and Sunstein, C. R. (2003). Libertarian paternalism. *American economic review*, 93(2):175–179.

Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

Weinstein, N. D. and Klein, W. M. (1996). Unrealistic optimism: Present and future. *Journal of Social and Clinical Psychology*, 15(1):1–8.

Appendix

General Instructions

Welcome!
You are about to take part in an experiment funded by several foundations for research purposes.

It is prohibited to communicate with the other participants during the experiment. Should you have any questions please ask the assistant. If you violate this rule, we shall have to exclude you from the experiment and from all payments.

During the experiment you will have the opportunity to make choices that will affect your earnings.
Your choices will not be communicated to the other participants. Anonymity will be preserved during and after the experiment: all the money you earn will be paid when the experiment is finished. Participation in the experiment will guarantee a minimum earning of 3 euro (show-up fee).

The experiment is composed by two parts. We start reading the instructions of Part 1. You will then receive the instructions for Part 2 in due course. In each Part, you may earn some euros. The earnings of each Part are independent, and you will be paid the sum of your earnings in every part.

Part 1

A grid with 100 boxes will appear on your computer.

Your task is to choose how many boxes to collect. So, you will be asked to choose a number between 1 and 100. Boxes will be collected in numerical order, starting from the box in the top left corner of the grid.

Every box collected is worth 0,03 euro. However, this earning is only potential since one of this boxes hides a bomb. You do not know where this bomb is. You only know that the bomb may be in any box with equal probability.

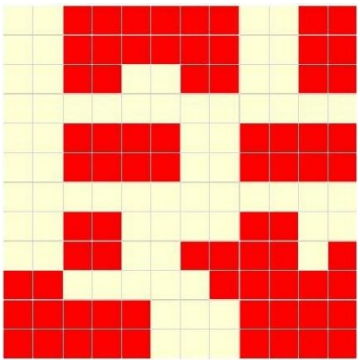
After choices have been made and confirmed, the computer will randomly determine which box contains the bomb. This random draw is made at the individual level, thus the box with the bomb can be different for every participant.

If the bomb is located in a box that you did not collect – i.e. the number of boxes you chose is smaller than the number of the box containing the bomb – you will earn 0,03 euro for each box collected.

In constrast, if you happen to collect the box where the bomb is located – i.e. if the number of boxes you chose is greater than or equal to the number of the box containing the bomb – the bomb will explode and destroy the earnings: thus, you will earn zero in Part 1.

Part 2

This part is made of 8 rounds: in each round you will see a figure on your screen, called *Target*, constituted by a grid of 12x12 squares, *called pixels*, colored either red or beige. The following figure is an example of a possible *Target*.



Example of a Target figure

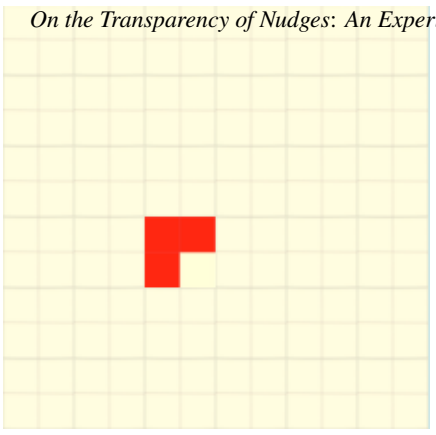
In each round you will have to replicate faithfully the *Target* figure, placing some blocks of 4x4 pixels (red or beige) in a 12x12 empty grid, called *Build*. The following figures include an example of a *Build* grid and some examples of 4x4 blocks.

Example of *Build* grid

Examples of 4x4 blocks

Each block may be placed in the *Build* grid by selecting a slot in the grid itself (the slots have the same size as a block, so there are 36 available slots) and clicking next on the block you want to insert. An inserted block may be replaced with another block whenever you want, repeating the same procedure just described. The next figure shows an example of selected slot (left) and inserted block (right).

Build **grid with selected slot**



Build **grid with inserted block**

In every round, the *Target* figure must be replicated with a margin of error of maximum 5 pixel: the round will not be considered successfully finished until more than 5 pixels in the *Build* grid differ from the figure in the *Target*.

Although there are 8 rounds, and hence 8 *Target* figures to replicate, before you begin the first round you will have to declare how many *Targets* you will be able to replicate successfully, knowing that the time to replicate the figure will diminish in each round. The following table shows the time that is available in each round.

Round	1	2	3	4	5	6	7	8
Available time	3 min. 0 sec.	2 min. 45 sec.	2 min. 30 sec.	2 min. 15 sec.	2 min. 0 sec.	1 min. 45 sec.	1 min. 30 sec.	1 min. 15 sec.

To simplify, we will call the number you will declare, the *Objective*. The *Objective* you will decide to state is very important because your final earnings will depend on it.

Your earnings

Your final earnings depend mainly from your capacity to replicate correctly at least as many *Target* figures as the *Objective* you have chosen. If you will not replicate at least as many *Target* figures as your *Objective*, your earning will be zero.

On the contrary, if you manage to attain your *Objective* you will earn 2,00 euro for every *Target* figure you will replicate, until you have reached your *Objective*.

Notice: once the *Objective* has been attained (if it doesn't coincide with the last round) the experiment will not end but you will have the opportunity to continue to replicate *Target* figures earning 50 cents for every extra figure you will correctly replicate. If you do not replicate a *Target* figure correctly in a subsequent round to your *Objective*, you will receive in any case the earnings that you have accumulated up until that point.

The following table summarizes the potential earnings according to the *Objective* (in columns) and the number of *Targets* correctly replicated by a subject (in rows).

num. of replicated *Target* figures	num. of declared *Target* figures (*Objective*)							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	2,00€	0,00€	0,00€	0,00€	0,00€	0,00€	0,00€	0,00€
2	2,50€	4,00€	0,00€	0,00€	0,00€	0,00€	0,00€	0,00€
3	3,00€	4,50€	6,00€	0,00€	0,00€	0,00€	0,00€	0,00€
4	3,50€	5,00€	6,50€	8,00€	0,00€	0,00€	0,00€	0,00€
5	4,00€	5,50€	7,00€	8,50€	10,00€	0,00€	0,00€	0,00€
6	4,50€	6,00€	7,50€	9,00€	10,50€	12,00€	0,00€	0,00€
7	5,00€	6,50€	8,00€	9,50€	11,00€	12,50€	14,00€	0,00€
8	5,50€	7,00€	8,50€	10,00€	11,50€	13,00€	14,50€	16,00€

You will thus receive the monetary earnings of this part, only if you will be able to replicate at least the number of *Target* figures you have initially declared. The residual seconds in each round *cannot* be used to increase the available time in the next round.

Notice that the final earnings will *not* depend on the amount of time you use in a particular round, but only on the capacity to replicate at least as many *Target* figures as those declared as *Objective*.

You will be able to go to the next round only when the *Target* figure in the current round will be replicated with a margin of error of maximum 5 pixels: in this case a "Proceed" button will appear on screen which will allow you to go to the next round.

End of experiment

The experiment will end in one of the following cases:

1. You have not replicated a *Target* figure in the available time, in a round that precedes your *Objective*. In this case your earnings will be equal to zero.
2. You have reached the *Objective* and you have not replicated a *Target* figure in a successive round. In this case your earnings will amount to 2.00 euro multiplied by your *Objective*, plus 50 cents for every *Target* figure you have solved beyond your *Objective* until the end of the experiment.
3. You have reached the final round (your earnings in this case will be calculated as in point two).

The experiment will proceed as follows:

1. You will have to answer some control questions on the experimental task (if you have doubts, raise your hand and wait for an assistant)
2. You will be asked to declare how many of the 8 *Target* figures you will want to replicate (from a minimum of 1 to a maximum of 8)
3. The first round of the experiment will begin and you will continue to replicate *Target* figures until you finish the experiment (for one of the three reasons listed above)
4. You will answer the final questionnaire
5. You will receive your final earnings.

After answering the control questions and before stating your Objective in order to familiarize with the task, you will have a few minutes to try to replicate a *Target* figure. This trial will not be paid: the purpose is to understand where the blocks can be inserted (slots), how to place them and how to replace them.
At the end of the trial we will answer any questions you may have about the task.