Le Maux, Benoît; Necker, Sarah

Working Paper

Honesty Nudges: Effect Varies with Content but Not with Timing

CESifo Working Paper, No. 10221

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Le Maux, Benoît; Necker, Sarah (2023) : Honesty Nudges: Effect Varies with Content but Not with Timing, CESifo Working Paper, No. 10221, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/271865

CESifo WORKING PAPERS

10221
2023

January 2023

Honesty Nudges: Effect Varies with Content but Not with Timing

Benoît Le Maux, Sarah Necker

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 10221*

Honesty Nudges: Effect Varies with Content but Not with Timing

Abstract

We use a ten-round online mind game to determine whether the effect of honesty nudges depends on timing and content. Reminding individuals about the right thing to do increases honesty. Including information that it is possible to assess an individual's dishonesty strengthens the effect of the intervention. Both types of intervention are similarly effective when they take place before an individual has made any decision or after individuals have played five rounds of the mind game. Nudging an individual after they have made five decisions allows us to add personalized information based on the individual's previous response; however, this does not increase honesty. Examining the reaction to nudges based on previous behavior shows that (presumably) honest and dishonest individuals respond by reducing overreporting. The effect of the different nudge content is driven by those previously dishonest.

JEL-Codes: C910, C920, M520, J280, J330.

Keywords: dishonesty, lying, cheating, honesty nudge, moral reminder, deterrence.

Benoît Le Maux
University of Rennes 1
CREM-CNRS, Condorcet Center
Faculté des Sciences Economiques
7 Place Hoche
France – 35065 Rennes Cedex
benoit.le-maux@univ-rennes1.fr

*Sarah Necker**
ifo Institute – Ludwig Erhard ifo Center for Social Market Economy and Institutional Economics
Gartenstraße 6
Germany – 90762 Fürth
necker@ifo.de

*corresponding author

We gratefully acknowledge financial support from the Fritz-Thyssen-Foundation, which had no involvement in the conduct of the research and preparation of the article. Declarations of interest: none. This study is registered in the AEA RCT Registry: ID "AEARCTR-0006416": the link to the pre-registriation is https://doi.org/10.1257/rct.6416. We thank Zareh Asatryan, Ulrich Glogowski, Marie Claire Villeval, Katharina Pfeil, Yilong Xu, and participants of the Lueneburg Workshop on Microeconomics 2022 and the Journees de Microeconomie Appliquee (JMA) 2022 (in Rennes) for valuable comments.

1. Introduction

Combatting unethical behavior is an objective of many organizations, including private companies, government agencies, schools, and academic institutions. While the standard economic theory of crime suggests that monitoring and punishment are effective strategies for achieving this goal (e.g., Becker, 1968), several recent studies examine whether simpler and cheaper strategies, such as honesty nudges, could be successful in preventing dishonesty. These nudges neither mandate honesty or forbid dishonesty, nor do they change the monetary incentives associated with different choices. Previous evidence suggests that the effect of nudges depends on the content and the context. While some studies find that honesty nudges successfully decrease cheating (e.g., Dunaiev and Khadjavi, 2021), others show that the effect is conditional, with nudges only reducing extreme lying (e.g., Jacquemet et al., 2021; Heinicke et al., 2019) or only working in a loaded environment (Jacquemet et al., 2019). There are also studies that find no effect from nudges (e.g., Dimant et al., 2020) or that nudges can even backfire by encouraging cheating (Zhao et al., 2019; Cagala et al., 2021). We contribute to the literature by examining whether the impact of nudges is dependent on whether individuals already made some choices in a sequence of moral decisions, whether the content of the nudges matters, and whether the timing and the content of nudges interact.

Our study is motivated by the observation that in many settings, individuals are nudged after they have already made decisions and possibly gave in to the temptation to cheat. Examples include taxpayers submitting annual tax declarations or researchers providing ethics declarations when submitting a paper – that is, after much of the work has been done. In these situations, nudges may interact with past behaviors. For example, if the individuals who cheated feel guilty, they may react more strongly to nudges than those who have not made any decision yet. However, most related studies focus on nudging before individuals make any decision. An exception is Kristal et al. (2020), who show that veracity statements are not any more effective when they are signed at the beginning of a self-report than at the end.

We also examine if honesty nudges are more effective when, in addition to reminding individuals about the right thing to do, they stress the possibility of monitoring. Deterrence nudges may change perceived incentives to be honest, yet, they do not actually change incentives: they may affect expected pecuniary costs, as well as reputational concerns. Antinyan and Asatryan (2019) show in their meta-analysis that tax compliance is unaffected by non-deterrence nudges that point to tax morale, but increased by deterrence nudges that reference auditing and sanctions.[1] By varying the timing, we can examine whether stressing the possibility of monitoring is more threatening to those who already cheated. Nudging individuals after they have made a decision implies that it is possible to refer to previous behavior. We examine if including personalized information based on past behavior has an effect, as it has been argued that nudges are more effective when they are individualized (e.g., Mills, 2020).

[1]To our knowledge, only two published studies systematically vary the content of honesty nudges in settings comparable to ours. Bryan et al. (2013) show that a nudge highlighting the implications of cheating for the actor's identity ("don't be a cheater") has a stronger effect than language focusing on the action ("don't cheat"). Dimant et al. (2020) vary the information provided on what others did, or what others believed one should do, as well as the framing (minority dishonesty/majority honest); they find that none of the nudges has an effect.

We study the impact of honesty nudges in a high-powered, online experiment.[2] In our study, 1,744 individuals participate in ten rounds of a wheel of fortune game. In each round, participants spin a wheel that shows six letters. The task is to guess the outcome of the wheel spin and to report whether their guess was correct. By comparing reported outcomes with chance, we can examine the effect of nudges on dishonesty at the aggregate level. In addition, we can assess dishonesty at the individual level by aggregating reports across multiple rounds.

Our between-subjects design consists of six treatments. Baseline treatment B contains no nudge and serves as a control group. In treatment Nudge M, we implement a moral reminder that repeats the instructions and stresses the importance of being honest. In Nudge MD, we complement Nudge M with a deterrence component that does not change the incentives, but stresses that we will compare reported outcomes to chance (moral reminder + deterrence). We implement these two nudges either before the first decision (R1) or before the sixth decision (R6). For Nudge MDI we provide subjects with additional information about whether we determined they followed the instructions (moral reminder + deterrence + individualized information). By design, this nudge can only be implemented in R6. In all treatments, the subjects are told that by continuing, they declare that they will honestly report their guesses.

We use a simple theoretical framework to explain how the nudges might affect the decision to lie in a dynamic setting. In our model, the evolution of dishonesty is not only affected by the nudges but also by past decisions. As a result, the agents' reaction to nudges may vary with their past behavior. This may in return affect the effects of nudges through time.

We find that all honesty nudges significantly decrease cheating. The moral reminder reduces the number of reported correct guesses by 13% when implemented before the first decision and by 21% before the sixth decision. Additionally stressing the possibility of monitoring reduces dishonesty by 27% in R1 and by 31% in R6. Hence, in line with previous literature, deterrence significantly increases the effectiveness of nudges. However, including individualized information has no significant additional effect, with reported correct guesses decreasing by 36%. The results rather suggest that Nudge MD is already perceived as having an individual component. The finding that the effect is quantitatively higher in R6 than in R1 aligns with the idea that guilt causes liars to react more strongly to nudges implemented after decisions have been made (in R6). However, the differences are at most significant at the 10% level. Our results suggest that in settings where the effect of nudge interventions does not deteriorate over time, as in ours, they should be implemented as early as possible.

We examine if the reaction to nudges implemented in R6 varies with past behavior, thus contributing to a recent literature studying how different types of cheaters react to honesty nudges (Heinicke et al., 2019; Jacquemet et al., 2021). We find that those (presumably) honest as well as those dishonest reduce over-reporting. The effect is more pronounced among those previously dishonest and their reaction is responsible for the effect of the nudges' content. Our suggestive evidence shows that those who suffer less from guilt/shame react more strongly to nudges. This might be taken as evidence that the honest types incur an infinite cost of lying while the dishonest types face a finite cost, as discussed in Kajackaite and Gneezy (2017).

Nudges are often the only possible intervention if cheating is not observable or if monitor-

[2]Previous studies are often based on a small number of observations. Two studies could not be replicated with larger samples, stressing the importance of high power (Verschuere et al., 2018; Kristal et al., 2020).

ing is very costly. We show that highlighting the possibility of comparing reported outcomes to chance is an effective means for reducing cheating in such settings. There are several possible applications. For instance, with regard to exams, some forms of cheating leave traces in the data that allow for inference on cheating (e.g., Lin and Levitt, 2020; Cagala et al., 2021). This is particularly relevant with regard to online exams, where the options to monitor cheating are very limited and cheating is thus very likely to occur. Yet, statistical models allow for indirect evidence of cheating (Cleophas et al., 2021; Bilen and Matros, 2021). Similarly, in order to overcome the challenge of detecting cheating in scientific articles, the results have been compared to the probability distributions, using for example Benford's Law (e.g., Tödter, 2009; Hüllemann et al., 2017). Another possible application is preventing tax evasion. In settings where information from other sources about possible tax fraud is unavailable, unsupervised machine learning can be used to identify outliers (e.g., Savić et al., 2022). Our evidence suggests that informing students, researchers, or taxpayers about the possibility of comparing their behavior to statistical distributions might reduce their cheating.

The remainder of this paper is structured as follows. Section 2 describes the experimental design. Section 3 develops a two-state Markov model that addresses the dynamics of cheating and discusses the tested hypotheses from a microeconomic perspective. In Section 4, we present the results from the experiment. Section 5 provides a discussion and the conclusion.

2. Experimental Design

2.1. Sample

The study was implemented online using Amazon Mechanical Turk (MTurk), an online crowdsourcing platform that connects employers with workers to perform tasks. These so-called Human Intelligence Tasks (HIT) are posted on the platform for potential workers to search and complete. Workers can complete any HIT for the amount of monetary compensation offered for the task. The nature of the platform allows for the collection of a large number of observations within a short time period and at low cost. This enabled us to conduct a high-powered study, which is important for increasing the reliability of the results (see, e.g., DellaVigna and Pope, 2018; Kristal et al., 2020). We recruited and paid participants using Cloudresearch. Mturkers who accepted the HIT were redirected to a Qualtrics survey.

Mturk is increasingly being used to run experimental studies in economics, many of which have been published in the discipline's top journals (e.g., DellaVigna and Pope, 2018; Exley and Kessler, 2022). However, concerns emerged in 2008 regarding the quality of MTurk data. In response to these concerns, several best practices were developed (e.g., Kennedy et al., 2020; Chmielewski and Kucker, 2020). Following this development, we include several provisions to ensure the quality of our data. We only accept workers with an approval rating of 95% and who have completed at least 500 HITs. We use Cloudresearch's feature to exclude low-quality participants, and Qualtric's feature to prevent multiple submissions from an individual.[3] Before the experiment starts, subjects are required to pass a bot control

[3] https://www.cloudresearch.com/resources/blog/new-tools-improve-research-data-quality-mturk/ and https://www.qualtrics.com/support/survey-platform/survey-module/survey

(captcha). To make sure that the subjects are paying attention, we include a screener question after the welcome page. The first three sentences suggest that the participants are being asked about how they were feeling. The question then explains that we want to know if the participants are actually taking time to read the instructions and ask them to ignore the question on how they are feeling; instead participants are asked to check only a certain response as their answer (following, e.g., Berinsky et al., 2014). Participants who indicate any other response are redirected to the end of the survey.

2.2. Contents of the questionnaire and the cheating game

The questionnaire consists of two tasks and a brief survey, as explained on the welcome page. Individuals are thanked for their participation and prompted for their informed consent to participate in an academic study "about individuals' memory and behavior in situations marked by randomness." The participants are also informed that they will receive a fee of $1 (conditional upon their answering all control questions correctly and completing the study), and that they are able to win a bonus payment of up to $2, based on their decisions and chance. The wording of the welcome page and the questionnaire is presented in Appendix A.1.

Before participants play the cheating game, they engage in three rounds of a memory game which has two purposes. First, the aim is to focus people's attention on memorizing letters. Since the cheating game is also about memorizing letters, we hope to distract participants from the main purpose of the study.[4] Second, the results allow us to assess if memory skills can explain the outcome of the cheating game. In each round, participants are shown a sequence of letters of increasing lengths (5, 7, and 9 letters) for ten seconds and asked to remember the letters. When the time is up, participants are asked to enter as many letters as they can remember. Entering the correct letters is neither conditional for continuing nor incentivized.[5]

Our cheating game is a modified version of the wheel of fortune cheating task proposed by Gsottbauer et al. (2022). Our wheel shows the first six letters of the alphabet from A to F. For ten rounds, the individuals must guess which letter the wheel of fortune will show and report if their guess was correct. In each round, individuals who report "Yes" (their guess and selected letter were the same) receive $0.20 while those who report "No" (their guess and selected letter were not the same) receive $0. Thus, there is a financial incentive to dishonestly report "Yes". To ensure that the participants understand the incentive scheme, we include two

`-options/survey-protection/`.

[4]This has two motivations. First, due to its short duration, in a large number of studies, the cheating game was implemented at the end of another study (Fischbacher and Föllmi-Heusi, 2013; Kajackaite and Gneezy, 2017). For reasons of comparability, including a second game allows us to get closer to these settings. Second, we aim to reduce experimenter demand effects. The use of non-deceptive obfuscation, e.g., employing filler tasks, has been proposed as a solution (Zizzo, 2010). Several related studies also aim to obfuscate the cheating game by framing it as a prediction game (e.g., Shalvi and De Dreu, 2014; Garbarino et al., 2019; Barfort et al., 2019).

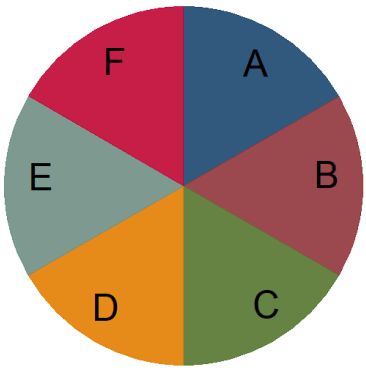
[5]The results from the memory game are as follows: 969 subjects enter three sequences correctly, 605 enter two sequences correctly, 149 enter one sequence correctly, and 21 subjects do not enter any sequence correctly. Those that enter all three sequences correctly are more likely to report that they guessed correctly than those that make one or more mistakes (3.8 vs. 4.6 reported correct guesses across the ten rounds, p-value: 0.000). This suggests that lack of memory cannot explain the high number of correct guesses reported in Section 4.1. On the contrary, it seems as if the more effort the subjects put into answering, the more likely they are to cheat.

control questions before the ten rounds start. Only subjects who correctly answer how much they can earn when they report "Yes" ($0.20) or "No" ($0) are allowed to continue.

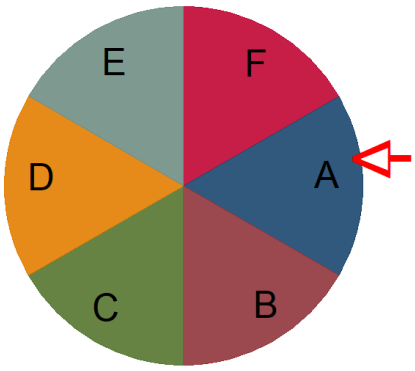
Each round proceeds as follows. Individuals are shown the wheel (see Figure 1a) and told that they have ten seconds to guess which letter the wheel will show and to remember their guess. Participants are automatically forwarded to the next screen after ten seconds, if they do not manually advance the screen themselves. The following screen shows the randomly selected letter of the virtual wheel, indicated by a red arrow (see Figure 1b). On this screen, participants report if the letter they guessed matches the letter selected by the wheel. Our game is a so-called mind game, where the individuals know that their lie cannot be exposed (e.g., Jiang, 2013; Kajackaite and Gneezy, 2017). As is common in the literature, we can infer group-level dishonesty by comparing the fraction/number of correct guesses reported by participants with the expected fraction/number of correct guesses if all subjects truthfully report what they observe. In addition, as described in Sections 2.3 and 4.3, we are able to estimate dishonesty at the individual level, as those with multiple "Yes" reponses can reasonably be suspected of dishonesty.

Figure 1: Wheel of Fortune

(a) Before spin



(b) After spin (example)



After the two games, participants complete a follow-up survey that asks about the individual's age, gender, education, income, hours per week spent working online, and level of affectedness by the COVID-19 pandemic (adopted from Belot et al., 2020). We also solicit self-assessed risk tolerance, shame, and guilt (adopted from Cohen et al., 2011), with the expectation that these factors might influence behavior. Finally, we ask participants to what extent they felt observed during the study. After completing all steps of the experiment, participants receive a completion code that allows them to collect their payment from Mturk.

2.3. Treatments

Individuals are randomly assigned to one of six treatments, in which we vary the timing and content of the honesty nudges in a between-subjects design. Our expectations with regard to the treatments are described in Section 3. The wording of the nudges is shown in Table 1.

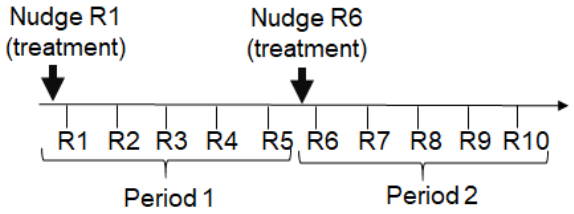
Table 1: Wording of Honesty Nudges (Treatments)

Treatment(s)	Wording
No Nudge B (baseline)	-
Nudge M: Moral reminder in either round R1 or R6	**Please be honest when reporting your guesses!** We expect the participants of our study to follow the instructions. Remember that we ask you to report if the letter you guessed is equal to the letter selected by the wheel. By continuing, you declare that you will honestly report your guesses.
Nudge MD: Moral reminder + deterrence message in either round R1 or R6	**Please be honest when reporting your guesses!** We expect the participants of our study to follow the instructions. Remember that we ask you to report if the letter you guessed is equal to the letter selected by the wheel. We (will) compare the number of guesses that you reported as correct to the statistical likelihood of obtaining that outcome. This allows (will allow) us to assess if you are following our instructions. By continuing, you declare that you will honestly report your guesses.
Nudge MDI: Moral reminder + deterrence message + individualized information in round R6 only	*Shown to those with 3, 4, or 5 correct guesses in period 1* **Please be honest when reporting your guesses!** We expect the participants of our study to follow the instructions. Remember that we ask you to report if the letter you guessed is equal to the letter selected by the wheel. We compare the number of guesses that you reported to be correct to the statistical likelihood of obtaining that outcome. You have reported that you guessed correctly in k of 5 rounds. Your reports suggest that you are NOT following our instructions. By continuing, you declare that you will honestly report your guesses. *Shown to those with 0, 1, or 2 correct guesses in period 1* **Thank you for being honest when reporting your guesses!** We expect the participants of our study to follow the instructions. Remember that we ask you to report if the letter you guessed is equal to the letter selected by the wheel. We compare the number of guesses that you reported to be correct to the statistical likelihood of obtaining that outcome. You have reported that you guessed correctly in k of 5 rounds. Your reports suggest that you are following our instructions. By continuing, you declare that you will honestly report your guesses.

Note: Bold and underlined letters are as shown in the experiment.

The baseline treatment No Nudge does not contain a message. In the two treatments Nudge M R1 and R6, subjects receive the following moral reminder: *Please be honest when reporting your guesses! We expect the participants of our study to follow the instructions. Remember that we ask you to report if the letter you guessed is equal to the letter selected by the wheel. By continuing you declare that you will honestly report your guesses.* The last sentence requires individuals to make a promise. According to psychological commitment theory, a promise has a binding function as individuals have a need to behave consistently (e.g., Kiesler, 1971). We made agreeing with the statement mandatory, as it is unclear how to deal with subjects who decline (see also, e.g., Heinicke et al., 2019). In addition, previous literature shows that almost all indidivuals agree if this is optional (e.g., Jacquemet et al., 2019, 2021). While in treatment Nudge M in R1 the message is shown on a separate screen after the wheel game instructions and before the first round, for Nudge M in R6, the message is shown before subjects enter the sixth round (i.e., after five decisions), as shown in Figure 2.

Figure 2: Flow of Wheel Task



In the treatment Nudge MD in R1, we complement the moral reminder with a deterrence component by adding the following two sentences: *We will compare the number of guesses that you report to be correct with the statistical likelihood of obtaining that outcome. This will allow us to assess if you are following our instructions.* In the treatment Nudge MD in R6, we add the same sentences in present tense. Deterrence messages used in letters to taxpayers stress the risk of detection (e.g., Dwenger et al., 2016; Bott et al., 2020) or emphasize the potential consequences of non-compliance (see, e.g., Fellner et al., 2013; Holz et al., 2020). To our knowledge, we are the first to exploit the possibility of making comparisons to chance to deter cheating (see Sections 1 and 5 for applications). We refrain from stressing potential consequences, as there are no sanctions. However, participants may be concerned that we will not approve their work, which could lead to a decrease in their Mturk approval rating.

Treatment Nudge MDI in R6 adds individualized information based on behavior in the first five rounds. In addition to the information contained in Nudge MD, we tell individuals how many times they reported guessing correctly in rounds one to five and – depending on the number of reported correct guesses – inform them whether their reports suggest irregularities. Those who report that they guessed correctly zero, one, or two times (probabilities 0.4, 0.4 and 0.16) receive a nudge thanking them for their honesty with the message: *Your reports suggest that you are following our instructions.* Those who report that they guessed correctly three, four, or five times (probabilities 0.03, 0.003 and 0.0001) receive a nudge asking them to be honest with the message: *Your reports suggest that you are NOT following our instructions.*

The decision to split the sample as described is based on the probabilities of the outcomes. As 3% or less is very low, we expect that almost all subjects reporting such outcomes cheat.

2.4. Descriptive statistics on sample

Of the 2,094 subjects that start our survey, 1,992 pass the screener question; those who fail it are redirected to the end of the survey (see Section 2.1). Of the subjects that continue, 53 drop out during the memory game, and 187 answer at least one control question of the cheating game incorrectly and are redirected to the end of the survey (see Section 2.2), and eight drop out during or after the cheating game. This implies that our sample consists of 1,744 subjects.

The average time to complete the questionnaire is 8 minutes 45 seconds.[6] The average age of participants is 38.5 years, 43% are female, and on average subjects work 19.8h online per week, as shown in Table A.1. To test for balance of sample characteristics, we regress these variables, as well as two variables measuring income, two variables measuring affectedness by the Covid-crisis, and the number of correct sequences in the memory game on the set of treatment dummies and test whether the estimated coefficients of these dummies are all jointly zero. For none of the variables, the F-tests indicates joint significance.

3. Testable Hypotheses

3.1. General setting

In line with our experiment, our simple model deals with a wheel game where the chance to win is $\pi = 1/6$. In each round r, agent i must decide whether to lie ($d_{i,r} = 1$) or not ($d_{i,r} = 0$) about whether their guess about the outcome of the wheel was correct. Following previous literature (e.g., Gneezy et al., 2018; Gerlach et al., 2019; Garbarino et al., 2019), we assume that lucky agents have no incentive to cheat and will honestly report their outcomes ($d_{i,r} = 0$). When a guess is not correct, agent i may choose to lie ($d_{i,r} = 1$), i.e., dishonestly report "Yes", if the pecuniary advantages b from doing so are larger than the lying costs c:

$$b - c(N, d_{i,r-1}) > \delta_{i,r}. \tag{1}$$

Lying costs c are caused by three types of motivations: a direct cost from lying (e.g., people want to maintain a positive self-image), a reputational cost (people care about how honest they appear to others), and the influence of social norms and conformity (people feel less bad when others are lying too; for a discussion of these motivations see, e.g., Abeler et al., 2019). The effects of the nudges $N \in \{\mathrm{B}, \mathrm{M}, \mathrm{MD}, \mathrm{MDI}\}$ depend on how they activate these motivations, as discussed in Section 3.2.[7] Since we would like to assess the effects of N conditional on past decisions, c also depends on $d_{i,r-1}$, i.e., whether agent i lied in the previous round.[8]

[6] The time does not vary significantly across treatments. Although we estimated the survey to take 10 minutes, it has been found that Mturkers complete the survey in less time (Berinsky et al., 2012).

[7] Note that our nudges are unlikely to activate social norms and conformity mechanisms, as we do not provide any information about the behavior of others. We include this motivation for the sake of completeness.

[8] Our empirical analysis shows that those who reported a high number of "Yes" responses in the previous period are also more likely to report a high number of "Yes" responses in the next round, see, e.g., Figure 7, which

To determine the share of liars in a round, we use probability distributions. We assume that variations in behavior are due to randomness: $\delta_{i,r}$ stands for a noise term that is independently distributed across agents and rounds. Let F be the continuous and increasing cumulative distribution of $\delta_{i,r}$. The probability that the unlucky agents are dishonest is $F(b-c)$ and that they are honest is $1-F(b-c)$. Given that c is past-dependent, those probabilities depend on the state (honest or dishonest) attained in the previous round. As a result, the allocation of agents between honesty ($d_{i,r}=0$) and dishonesty ($d_{i,r}=1$) is described as a two-state Markov model with the following transition matrix:

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}, \tag{2}$$

$$\begin{aligned}
p_{00} &= \Pr(d_{i,r}=0|d_{i,r-1}=0) = \pi + (1-\pi)\times[1-F(b-c(N,0))], && (3)\\
p_{01} &= \Pr(d_{i,r}=1|d_{i,r-1}=0) = (1-\pi)\times F(b-c(N,0)), && (4)\\
p_{10} &= \Pr(d_{i,r}=0|d_{i,r-1}=1) = \pi + (1-\pi)\times[1-F(b-c(N,1))], && (5)\\
p_{11} &= \Pr(d_{i,r}=1|d_{i,r-1}=1) = (1-\pi)\times F(b-c(N,1)). && (6)
\end{aligned}$$

These conditional probabilities express the chances for agents to be dishonest (or not) in round r given their behavior in round $r-1$. By construction, $p_{00}+p_{01}=1$ and $p_{10}+p_{11}=1$.

When starting the game, since they did not have the possibility to cheat, all agents belong to the honesty state, i.e., $d_{i,0}=0$. Now, let $\mathbf{M}_r=(m_{r0},m_{r1})$ describe the number of honest and dishonest agents in round r, respectively, and let $\alpha_r = m_{r1}/(m_{r0}+m_{r1})$ represent the resulting share of liars. Since behaviors are past-dependent, this allocation of agents among the two Markov states can be expressed as $\mathbf{M}_r=\mathbf{M}_{r-1}\mathbf{P}$. In other words:

$$\alpha_r = \frac{p_{01}m_{r-1,0}+p_{11}m_{r-1,1}}{m_{r-1,0}+m_{r-1,1}} = p_{01}(1-\alpha_{r-1})+p_{11}\alpha_{r-1}. \tag{7}$$

There are two channels. First, lying is directly affected by changes in the transition matrix: if lying costs c increase, then, from equations 4 and 6, the conditional probabilities to be dishonest p_{01} and p_{11} decrease, reducing the fraction α_r of liars. Second, lying is assumed to be past-dependent: if $p_{11}>p_{01}$, a lower share of liars in round r may in turn reduce the share of liars in round $r+1$. Thus, a change of the share of liars can have long run effects.

Note that our empirical analysis focuses on the number of "Yes" responses in five rounds to ease the interpretation of the results (see Section 4).[9] Formally, the share of "Yes" responses is equal to $\pi+(1-\pi)\alpha_r$, and the average number of "Yes" responses, say for rounds R to $R+K$, can be specified as:

$$y_R = \sum_{r=R}^{R+K} \pi + (1-\pi)\alpha_r = (K+1)\pi + (1-\pi)\sum_{r=R}^{R+K}[p_{01}(1-\alpha_{r-1})+p_{11}\alpha_{r-1}]. \tag{8}$$

Thus, any change in p_{01} and p_{11} that affects the share of liars in rounds R to $R+K$ applies equivalently to the mean of Yes. In our study, R is either equal to 1 or 6, and K equal to 4.

suggests that liars have lower lying costs, i.e., $c(N,1)<c(N,0)$, and thus cheat multiple times.

[9] As shown in Figure A.1 in the Appendix, we find the same results when we only consider the first round after the nudges, or all the rounds individually (not averaged across five rounds) after the nudges.

3.2. Expectations

Effect of Nudge M. As long as a rule is ambiguous or not fully defined, as in the baseline treatment (No Nudge B), people can justify their actions by strategically perceiving the rules in their favor (Shalvi et al., 2015) or by ignoring their own moral standards (Mazar et al., 2008). We hypothesize that Nudge M will guide participant attention to their own standards of integrity, increasing self-image concerns and inducing direct lying costs (see also Mazar et al., 2008; Heinicke et al., 2019). In addition, it could be that Nudge M has a repetition effect, since we repeat part of the instructions (see, e.g., Bursztyn et al., 2019). To summarize, our expectations about the effect of Nudge M in rounds R to $R+K$ are as follows.

Hypothesis 1. *We have $c(M,d_{i,r-1}) > c(B,d_{i,r-1})$. As a result, y_R, the average number of "Yes" responses in rounds R to $R+K$, will be lower under Nudge M than with No Nudge B.*

Effect of Nudge MD. Since the wheel game is a mind game, participants might think that it is not possible to observe dishonesty. Nudge M does not contain any information regarding observability. However, with Nudge MD, we inform participants about a monitoring process. Although we do not threaten sanctions, this may affect the subjective probability of being detected, potentially causing concerns that we will not approve their work (see Section 2.3). In addition, the nudge could increase the cost of lying if agents care about their reputation. The nudge clearly highlights that reporting an unusually high number of successful rounds will be interpreted as lying. Reputational concerns have been described as a positive function of the probability that a statement is perceived as dishonest (e.g., Abeler et al., 2019; Gneezy et al., 2018; Khalmetski and Sliwka, 2019; Dufwenberg and Dufwenberg, 2018). For instance, subjects are more likely to lie when their outcomes cannot be observed by the experimenter than when the experimenter can later verify the actual outcome (Gneezy et al., 2018). Thus, we hypothesize that Nudge MD will be more effective in reducing cheating than Nudge M. Hence, in rounds R to $R+K$, Nudge MD is expected to be more effective than Nudge M:

Hypothesis 2. *We have $c(MD,d_{i,r-1}) > c(M,d_{i,r-1})$. As a result, y_R, the average number of "Yes" responses in rounds R to $R+K$, will be lower under Nudge MD than under Nudge M.*

Effect of Nudge MDI. We exploit the timing dimension to implement an individualized Nudge MDI in the middle of the game, which contains information that depends on past behavior. We inform participants how many times they reported guessing correctly in the last five rounds, making it more salient that we are able to observe behavior. In addition, we treat those who were presumably honest and those who were most likely dishonest differently. Those who report more than two "Yes" responses are informed that "*your reports suggest you are NOT following our instructions*." Those who report up to two "Yes" responses receive a message thanking them for "*following our instructions*." We expect that this information increases the expected probability of detection and sanctioning, as well as reputational concerns. If nudges backfire, in that those who were honest are made aware of the possibility of cheating or feel offended by a message asking them to "*please be honest*" (Zhao et al., 2019; Cagala et al., 2021), a nudge thanking them for their honesty could be more effective. Hence, in rounds $R=6$ to $R+K=10$, we expect Nudge MDI to be more effective than Nudge MD.

Hypothesis 3. *We have $c(MDI, d_{i,r-1}) > c(MD, d_{i,r-1})$. As a result, y_R, the average number of "Yes" responses for rounds R to $R+K$, will be lower under Nudge MDI than under Nudge MD.*

Different reactions. Past behaviors $d_{i,r-1}$ may interact with the nudge in the cost function inducing different reactions depending on whether a participant lied in the previous round. Let Δp_{01} and Δp_{11} denote the (negative) effects of a nudge $N^* \in \{\mathrm{M}, \mathrm{MD}\}$ compared to baseline B on the probability of lying when the agent was previously honest or dishonest. We have:

$$\Delta p_{01} = (1-\pi) \times [F(b - c(N^*, 0)) - F(b - c(B, 0))], \tag{9}$$

$$\Delta p_{11} = (1-\pi) \times [F(b - c(N^*, 1)) - F(b - c(B, 1))]. \tag{10}$$

There are three scenarios. First, the effect of nudges M and MD could be higher among those who lied previously, if they engage in moral cleansing or conscience accounting. The terms describe the act of behaving more morally after a past transgression to avoid guilt and to close the gap between one's desired and one's moral self-image (e.g., Ploner and Regner, 2013; Gneezy et al., 2014; West and Zhong, 2015). The finding that moral balancing is stronger when people are observed suggests that not only self-image but also reputational concerns matter (Rotella et al., 2019). In these cases, making people aware of their misdeeds should result in compensatory behavior (Ilies et al., 2013). We would have $\Delta p_{11} < \Delta p_{01} < 0$.

Second, the effect of nudges M and MD could be lower among liars. It has been shown that people adopt a wide range of strategies to avoid responsibility for past misdeeds and to maintain their self-image. Lying has been shown to result in moral disengagement, which describes the process of making detrimental conduct acceptable by persuading oneself that the transgression is actually permissible (e.g., Shu et al., 2011; Galeotti et al., 2020). To avoid having to update one's self-image, liars could pay less attention to nudges. Individuals might even "willfully ignore" evidence about the harmful impact of their decisions (e.g., Grossman and van der Weele, 2016). Additionally, it could be that individuals feel they have already signaled their dishonesty in previous rounds and, therefore, cannot rectify their reputation. For that reason, they might continue or even extend their lying. In those cases, we have $\Delta p_{01} < \Delta p_{11} < 0$. Third, it could be that $\Delta p_{01} = \Delta p_{11}$ and that the effect of N^* is independent of the previous decisions.

Hypothesis 4.

a. If $\Delta p_{11} < \Delta p_{01} < 0$, Nudges M and MD have a larger effect on liars.

b. If $\Delta p_{01} < \Delta p_{11} < 0$, Nudges M and MD have a lower effect on liars.

c. If $\Delta p_{11} = \Delta p_{01}$, the effect of Nudges M and MD is independent from past behavior.

Timing of Nudges. A central question to be answered is if timing modifies the effect of Nudges M and MD. From equations 8, 9 and 10, the effect of $N^* \in \{\mathrm{M}, \mathrm{MD}\}$ compared to baseline B in rounds R to $R+K$ can be written as:

$$\Delta y_R = (1-\pi) \sum_{r=R}^{R+K} [\Delta p_{01}(1-\alpha_{r-1}) + \Delta p_{11}\alpha_{r-1}]. \tag{11}$$

The average treatment effect depends on equations 9 and 10 and the previous shares of liars α_{r-1}, $r = R$ to $R+K$. For instance, if the share of liars is small, nudging in R6 may not be more effective even if liars respond more strongly to the nudge. In other words, a necessary (but not sufficient) condition for observing a varying effect of nudges across time is a significant difference in the share of liars at the time the nudge is sent. Formally:

Hypothesis 5. *If either H4.a or H4.b is true and if* $\sum_{r=R1}^{R1+4} \alpha_{r-1} \neq \sum_{r=R6}^{R6+4} \alpha_{r-1}$*, then the effect of Nudges M and MD differs across time.*

In summary, dishonesty may be affected in several ways, involving complex combinations of factors in a dynamic setting. Our simple theory illustrates those interactions with the aim of clarifying each particular channel. First, our discussion highlights that different contents can activate different types of motivations to be honest. Second, in our model, the evolution of dishonesty is not only affected by the nudges N but also by past behavior $d_{i,r-1}$. We consider this in the empirical analysis. In particular, estimating the transition matrix, and further examining whether reactions to nudges differ, allows us to disentangle those channels.

4. Experimental Results

4.1. Effects of Nudges on Average Number of Yes

Our main variables of interest are the average number of reported correct guesses in the first five rounds, denoted as period 1, and in the last five rounds, denoted as period 2. The expected number of correct guesses in a period is 0.8225. For these variables, the power to detect an effect size of 0.5 is close to 100% for a one-sample t-test (comparison with 0.8225) and 99% for a two-sample t-test (comparison of treatments). Conversely, assuming a power of 80%, we are able to target a minimum effect size of 0.166 and 0.233, respectively.[10]

As shown in Figure 3, in both periods and in all treatments, the reported number of correct guesses is significantly higher than the expected number (p-values 0.000).[11] For example, without a nudge, participants report on average 2.4 correct guesses in period 1 and 2.6 in period 2. Although cheating is prevalent, the averages are far from the maximum. In line with previous studies, despite our anonymous setting, which makes it impossible to prove that someone lied, a substantial fraction does not maximize the financial gain (e.g., Abeler et al., 2019). The results are the same when we examine the fraction of participants reporting a correct guess in each of the ten rounds, as shown in Figure A.1 in the Appendix.[12]

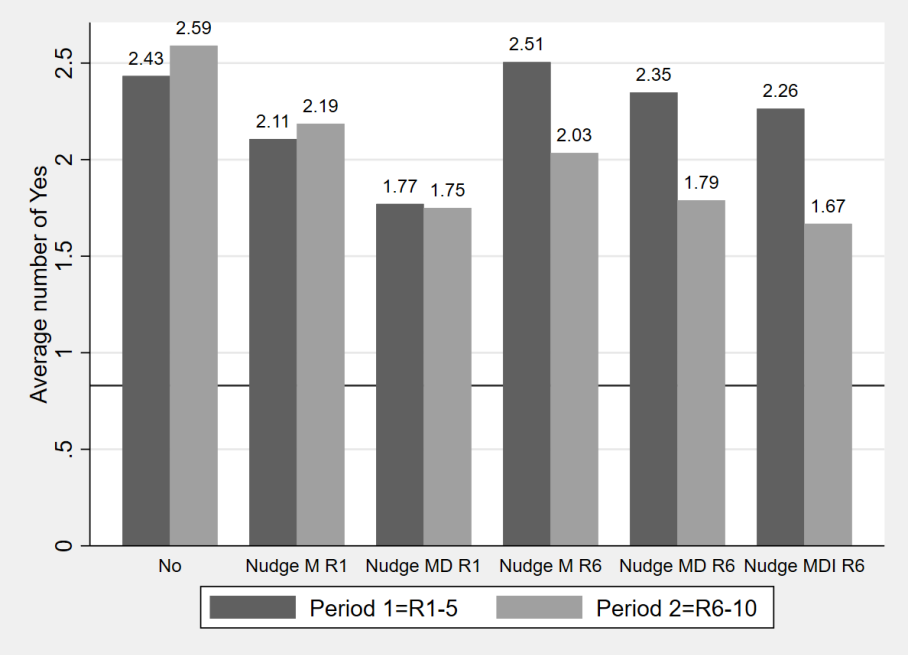
To assess the effect of nudges implemented at different points of time, we focus on the effect of the nudge in the first five rounds after the message was shown to participants. Hence,

[10] Note that the smallest sample sizes are 288 (Nudges MD-R1) and 290 (Nudges M-R1 and MD-R6). If we account for the largest possible standard deviation (i.e., period 1 of the baseline group, SD=1.669) those values represent differences in the average number of "Yes" responses that are equal to 0.27 and 0.388, respectively.

[11] Unless otherwise indicated, we report t-tests. The results are largely the same with non-parametric tests. In the text, we consider differences to be statistically significant when the significance level is less than 5%.

[12] The likelihood of guessing correctly in a round is 0.167. Across all rounds and treatments, the fraction of "Yes" responses is significantly higher. For example, in the first round of No Nudge, 47% of participants report guessing correctly. This implies a liar rate of 36% ((0.47-0.167)/(1-0.167)), see Figure A.2.

Figure 3: Reported Correct Guesses by Treatment and Period



Note: Presented are the number of times an individual reports guessing correctly in five rounds. The expected number of "Yes" responses is 0.8225, as indicated by the horizontal line. Nudge M is a moral reminder, Nudge MD adds a deterrent component to Nudge M, Nudge MDI adds individualized information to Nudge MD. R1 is round 1, R6 is round 6. The numbers of obs are as follows: No: 293, Nudge M R1: 290, Nudge MD R1: 288, Nudge M R6: 291, Nudge MD R6: 290, Nudge MDI R6: 292.

for nudges implemented in R1 we compare behavior in period 1 across control and nudge treatments. The same is done for nudges in R6 in period 2. To examine if the effectiveness of nudges varies with their timing, we hold the content constant and compare the difference in behavior between the baseline and the respective nudge treatments across R1 and R6 nudges.

First, we examine if R1 nudges are effective and whether the effect varies with the content. In line with Hypothesis 1, reminding subjects in R1 about the rules and asking them to report honestly (Nudge M) decreases the number of correct guesses in period 1 from 2.43 to 2.11, i.e., by 13% (p-value: 0.016), as shown in Figure 3. When we inform individuals that we are able to assess if their behavior deviates from chance (Nudge MD), the average number of "Yes" responses decreases to 1.77, i.e., a 27% (p-value: 0.000) decrease compared to the baseline. Both differences are statistically significant, suggesting that both nudges are effective. To examine if the deterrence nudge is more effective than the nudge that only contains the moral reminder, we obtain the empirical bootstrap distribution of the difference (500 replications, sampling with replacement) and assess if it is significantly different from zero. Accordingly, supporting Hypothesis 2, adding deterrence significantly increases the effectiveness of the nudge (p-value: 0.0097). Still a large fraction cheats, suggesting that many understand that there are no sanctions and we are unable to prove if someone was dishonest.

Next, we study the effectiveness of the different nudges in R6. In Figure 3, the bars depicting behavior in the five rounds after the nudge show that the number of reported cor-

rect guesses decreases from 2.59 in the baseline to 2.03 in Nudge M, i.e., a 21% decrease (supporting Hypothesis 1). Adding information about the possibility of monitoring in Nudge MD decreases the reported correct guesses to 1.79 (31% decrease). Further adding individual-specific information decreases the number of reported successes to 1.67 (36% decrease). The differences to the baseline treatment are highly significant (p-values: 0.000). Nudge MD is again more effective than Nudge M (p-value: 0.043), supporting Hypothesis 2. The same applies to the individualized Nudge MDI (p-value: 0.003). Yet, Nudge MDI is not significantly more effective than Nudge MD (p-value: 0.281), i.e., Hypothesis 3 is not supported. Hence, while making individuals aware of the possibility of assessing dishonesty decreases overreporting, on average it does not seem to be necessary to observe individual behavior.

We investigate if Nudges M and MD are more effective when they are implemented before (R1) or after (R6) individuals have made decisions (Hypothesis 5). Using bootstraping, we compare differences in behavior between the baseline and the R1 nudge treatment in the first five rounds to differences in behavior between the baseline and the R6 nudge treatment in the last five rounds (difference-in-difference). The decrease in reported correct guesses caused by the moral reminder (Nudge M) is 0.23 higher when individuals are nudged in R6 (p-value: 0.127). For the deterrence nudge, the decrease is 0.14 higher under the R6 nudge (p-value: 0.316). While the direction of the effect is in line with the idea that guilt causes individuals to react more strongly when nudges are provided after some decisions have been made, the differences are non-significant. Thus, we obtain no support for Hypothesis 5. We also examine if Nudge MDI in R6 is more effective than Nudge MD in R1. In line with the other results, the difference in the decrease is only slightly higher in the R6 nudge (p-value: 0.064).

A possible explanation of this finding is that the share of liars is rather stable in the baseline treatment (as shown in Figure A.1), which may counteract observing a significant effect of timing across periods, even if liars react more strongly to nudges (as found in Section 4.3).

4.2. Effect of Nudges on Results Per Round and on the Distribution of Reports

The described effects are also reflected in the results at the round level, as shown in Figure A.1 in the Appendix. While the nudges in R1 shift the fraction of "Yes" responses to a lower level in all rounds compared to the baseline (see Figure A.1 (a)), the nudges in R6 do so from round 6 onwards (see Figure A.1 (b)). The pattern of the results across treatments is the same as before, supporting Hypothesis 1 and 2, and also found when we only consider the first round after the nudges. The results at the round level also allow us to assess the effect on the share of liars. As shown in Figure A.2, for example, we find that Nudge M in R1 decreases the share of liars in Round 1 from 36% to 31%. Nudge M in R6 decreases the share of liars in round 6 from 44% to 28%. Although this example suggests that in line with Hypothesis 5 the effect of the nudging is stronger in R6 it should be noted that the difference is smaller in later rounds.

We also examine how our nudges affect the distribution of reported correct guesses in the five rounds after the nudge, as shown in Figure 4. Here we follow previous studies examining how nudges affect the fraction of partial liars (not lying to a maximum extent) and extreme liars (lying to a maximum extent) (Heinicke et al., 2019; Jacquemet et al., 2021). According to Kolmogorov-Smirnov tests, the distributions are different from the baseline treatment in all

nudge treatments except for Nudge M in R1 (all p-values<0.01). The pattern is very similar across treatments: while the fraction reporting zero, one, or two correct guesses increases, the fraction reporting three, four, or five decreases, compared to the baseline. This suggests that those reporting three, four, or five correct guesses lie, which is plausible and in line with our expectations, as the chances to obtain these outcomes are very low (see footnote of Figure 4). However, the fraction of those reporting having guessed correctly two times is in all treatments higher than the expected value of 0.16, suggesting that this fraction includes liars.

In analyzing whether the distributions differ across nudge contents, we find significant differences between Nudge M and Nudge MD at the 10% level (under R1 and R6 nudges) and between Nudge M and Nudge MDI at the 1% level (naturally only under the R6 nudge). In all nudge treatments, we observe the sharpest decrease among those reporting five correct guesses, i.e., of extreme liars, compared to the baseline. In the nudges implemented in R1, the difference across nudge contents is caused by a lower fraction reporting guessing correctly three, four, and five times in Nudge MD R1 than in Nudge M R1. In the R6 nudges, we observe that the decrease in extreme liars is similar in all nudges, while the fraction of those reporting three or four (partial liars) varies across nudge types. Concerning the timing (holding content constant), we do not find significant differences in the distributions.

4.3. Does the Reaction to Nudges Depend on Past Behavior?

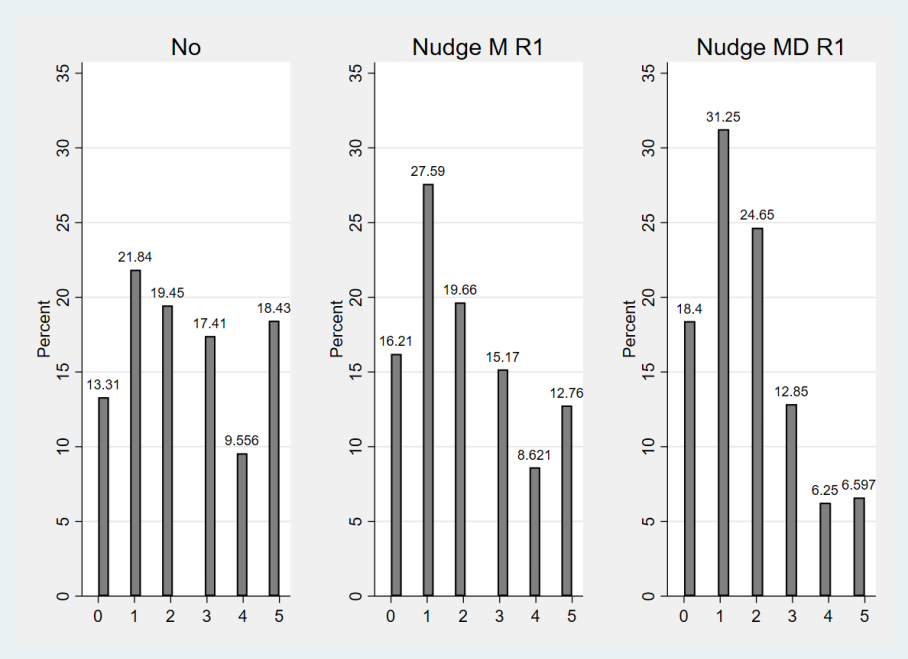
To consider the dynamic nature of individual decisions and the possibility of path-dependent behavior, our theory relies on transition matrices, i.e., the conditional probabilities to be dishonest or not in round r given behavior in round $r-1$. The matrices allow us to consider the possibility that the reaction to nudges varies with past decisions (Hypothesis 4). Since cheating cannot be observed with certainty at the individual and round level, we cannot directly estimate transition matrices. However, we can approximate them using the conditional probability of providing a "Yes" response given the report in the previous round.

In Figure 5, we examine the probability of reporting "Yes" based on having reported "Yes" (subfigure a and c) or "No" (subfigure b and d) in the previous round. As described in equations 9 and 10, since our main interest is in the change of probabilities compared to the baseline, we report the difference in the fraction of Yes compared to this treatment. The lower a curve is, the larger is the negative difference with respect to the baseline. Subfigure (a) suggests that Nudge MD R1, but not Nudge M R1, lowers the likelihood that someone reporting "Yes" in the previous round reports another "Yes" in the following round. The effect is rather stable across rounds. With regard to those who reported "No" in the previous round (subfigure b), we find that the probability to report "Yes" decreases with a nudge in later rounds. Regarding the nudges in R6 (subfigures c and d), we find a drop in the likelihood of reporting "Yes", in particular, in the rounds immediately after the nudge. Overall, Figure 5 suggests that the probabilities conditional on "No" are less affected. With the exception of Nudge M R1, the decrease is quantitatively higher among the "Yes" group than among the "No" group in almost all rounds, suggesting asymmetric reactions to nudges as stated in H4a.

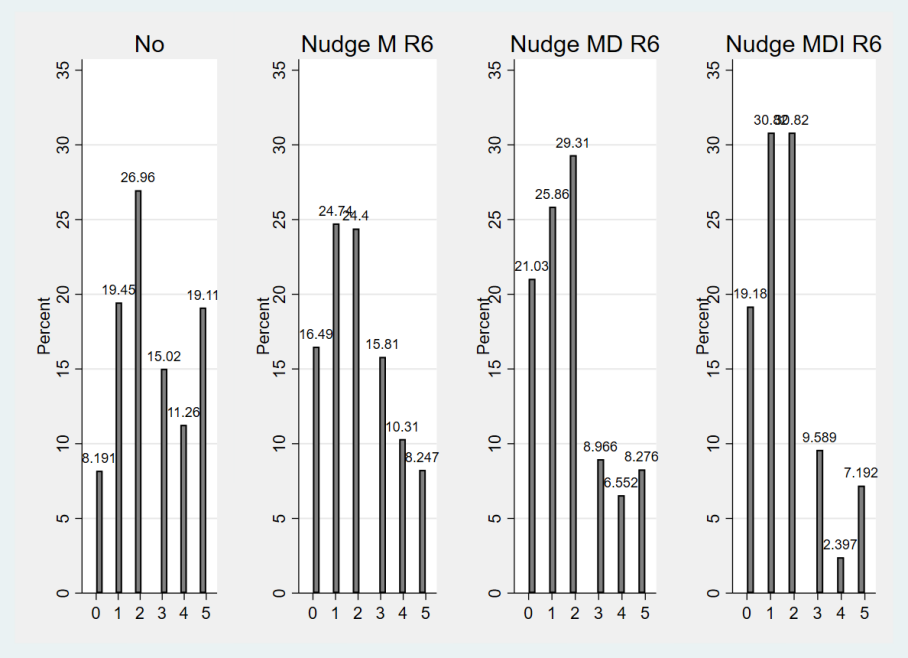
In Figure 6, we focus on the conditional probabilities of reporting "Yes" in R6 given the behavior in R5, which allows us to take a more detailed look at the effect of nudging

Figure 4: Distributions of Reported Correct Guesses by Treatment

(a) Effect of R1 Nudges on Reports in Period 1



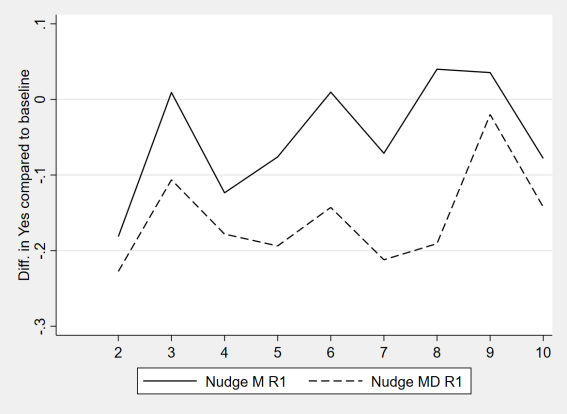
(b) Effect of R6 Nudges on Reports in Period 2



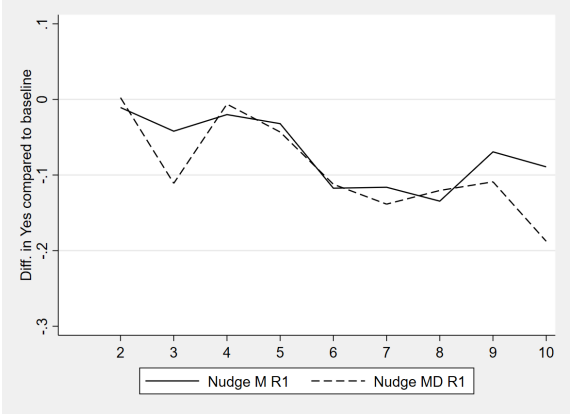
Note: Reported are the number of times an individual reports correctly guessing in five rounds. The expected likelihood to obtain those outcomes by chance (binomial distribution) are: 0.4 for zero "Yes", 0.4 for one "Yes", 0.16 for two "Yes", 0.03 for three "Yes", 0.003 for four "Yes", 0.0001 for five "Yes". Period 1 comprises rounds 1 to 5, period 2 rounds 6 to 10. Nudge M is a moral reminder, Nudge MD adds a deterrent component to nudge M, and Nudge MDI adds individualized information to Nudge MD. R stands for Round, R1 is round 1, R6 is round 6. The numbers of obs are as follows: No: 293, Nudge M R1: 290, Nudge MD R1: 288, Nudge M R6: 291, Nudge MD R6: 290, Nudge MDI R6: 292.

Figure 5: Transition Matrices

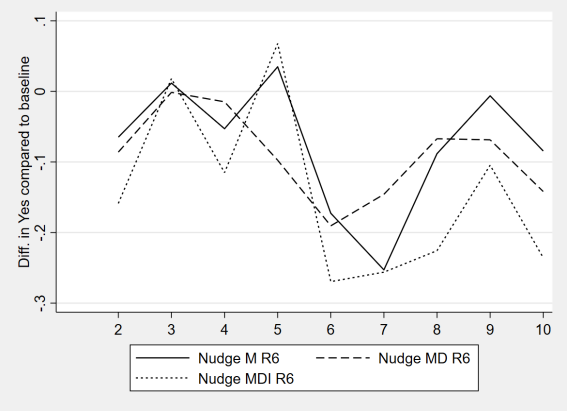
(a) From Yes to Yes (p_{11}) - R1 Nudges



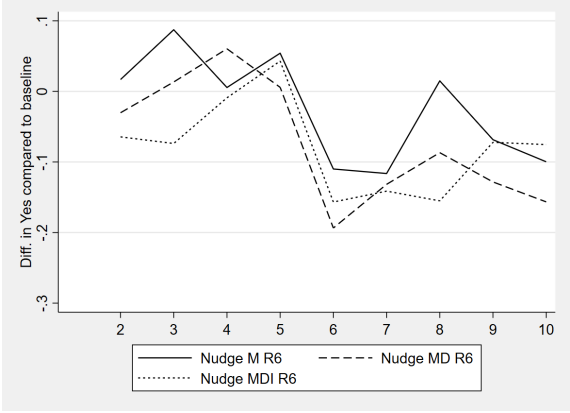
(b) From No to Yes (p_{01}) - R1 Nudges



(c) From Yes to Yes (p_{11}) - R6 Nudges



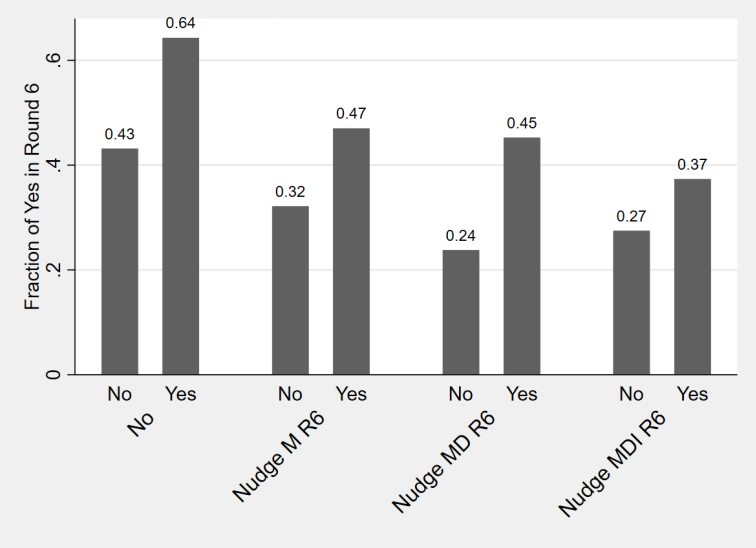
(d) From No to Yes (p_{01}) - R6 Nudges



Note: Depictions of the fractions of "Yes" responses based on the report in the previous round ("Yes"/"No") compared to the baseline. Section 3 describes the definition of p_{11} and p_{01}. Nudge M is a moral reminder, Nudge MD adds a deterrent component to nudge M, Nudge MDI adds individualized information to Nudge MD. R stands for Round, R1 is round 1, R6 is round 6.

holding past-dependency constant and test for the statistical difference of the effect across those reporting previously "Yes" vs. "No." Compared to the baseline, Nudge M decreases the likelihood of reporting "Yes" among those reporting "No" in R5 from 43 to 32%, i.e., by 11ppts (p: 0.053). Among those reporting "Yes" in R5, the decrease is 17ppts (p: 0.003). We use bootstrapping to examine if the decrease is more pronounced among the "Yes" than among the "No" group. The difference in the change of 6ppts is not significantly different from zero (p: 0.438). With regard to Nudge MD, both groups face the same decrease in the fraction of "Yes", i.e., a 19ppts decrease (p: 0.002); hence, the change is not statistically different across groups (p: 0.97). With regard to Nudge MDI, the decrease is 16ppts if the respondent previously reported "No" (p: 0.005) and 27ppts if "Yes" (p: 0.000). The difference between the two groups is again not statistically significant (p: 0.136). A similar effect emerges when we examine the difference between "Yes" and "No" in other rounds. Hence, although the effect is quantitatively stronger among those who reported "Yes" previously, in line with Hypothesis 4a, the difference is never significant.

Figure 6: Fraction of "Yes" in R6 Conditional on Report in R5



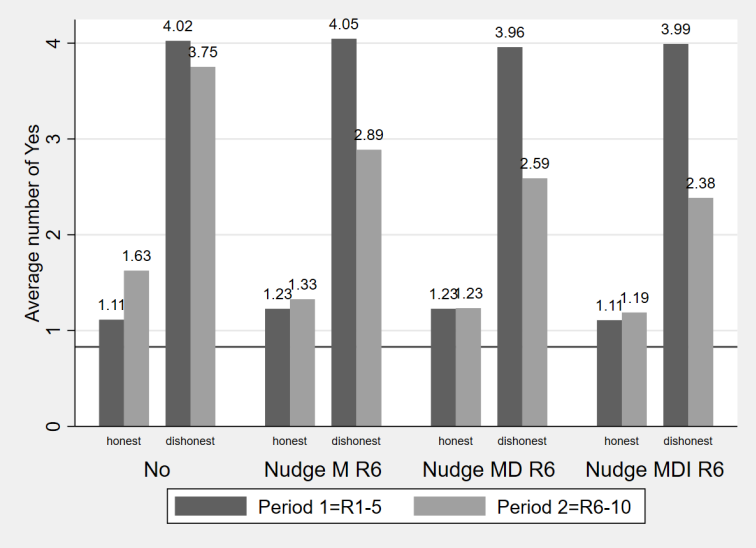
Note: Reported is the conditional probabilities of reporting "Yes" in round 6 given the reported outcome in round 5. Nudge M is a moral reminder, Nudge MD adds a deterrent component to nudge M, Nudge MDI adds individualized information to Nudge MD.

As another test of the interaction between nudges and past behavior, we use the nudges implemented in R6 and split the sample in two groups based on reports in the first five rounds, as in our treatment Nudge MDI (see Section 2.3) and as suggested by the distributions reported in Figure 4. The first group consists of participants reporting correctly guessing three, four, or five times in the first five rounds, i.e., those previously "dishonest." The second group consists of participants reporting correctly guessing zero, one, or two times; i.e., those "(presumably) honest." As described above, it is possible that not all of those who claimed two correct guesses were actually honest, as the fraction is higher than the expected value.[13] Importantly,

[13] As a robustness check, we examine behavior by the number of correct guesses in the first period in Figure A.3.

in the first period, the fraction of dishonest participants is not statistically different across the four treatments (Baseline/Nudge M R6: 45%, Nudge MD R6: 41%, Nudge MDI R6: 40%).

Figure 7: Do (Presumably) Honest React Differently than Dishonest People to R6 Nudges?



Note: Reported are the number of times an individual reports correctly guessing in five rounds. The expected number of "Yes" responses is 0.8225, as indicated by the horizontal line. (Presumably) honest participants report correctly guessing 0, 1, or 2 times in period 1, dishonest participants report correctly guessing 3, 4, or 5 times.

We analyze the behavior of these two types during the two periods, as shown in Figure 7. Figure 7 shows that in the baseline condition, those who were (presumably) honest increase the number of "Yes" responses by 0.5 (p-value: 0.000) while those who were dishonest decrease the number of "Yes" responses by 0.27 (p-value: 0.002), suggesting that without a nudge moral balancing takes place. However, the figure also shows that behavior is past-dependent, in that the number of "Yes" responses in the second period is substantially higher among those who were dishonest in the first period.

We compare the treatments including a nudge in R6 with the baseline treatment. For all treatments, a nudge lowers the number of correct guesses among those who were (presumably) honest (p-value ≤ 0.001) but there are no differences across the three nudges. Among the dishonest, the reported number of "Yes" responses is also significantly lower than in the baseline in all treatments. Dishonesty is slightly lower under Nudge MD than under Nudge M (p-value: 0.09), and significantly lower under Nudge MDI (p-value: 0.003). Yet, the difference between Nudge M and Nudge MDI is again not significant. We use bootstrapping to examine if the change among the honest is significantly lower than the change among the dishonest.

In the baseline treatment, participants who reported five "Yes" responses in period 1 report a lower number of "Yes" in period 2 on average; participants who reported one "Yes" in period 1 significantly increase the reports of "Yes" responses in period 2; no significant difference is observed for the other participants. In all nudge treatments, we find a significant decrease in the number of correctly reported guesses, except for the participants who reported one "Yes" in period 1.

We find that this is the case in all treatments (Nudge M p-value: 0.013, Nudge MD p-value: 0.002; Nudge MDI p-value: 0.000). Hence, a nudge prevents those who were (presumably) honest in the past from becoming dishonest, but more strongly reduces overreporting among the dishonest. We also examine whether this was caused by different attention paid to the nudges. We find that, although not significantly so, the dishonest participants tend to spend more time reading the nudges (see Table A.2).[14] The results suggest that moral cleansing but not willful ignorance matter in the reaction of liars to the R6 nudges. In addition, we find that the content of the nudge matters only for the dishonest, supporting Hypotheses 1 and 2 for this group. The patterns are unchanged when we categorize those reporting two correct guesses as liars, as shown in Figure A.4. The average number of "Yes" of those coded as "dishonest" decreases, while the average number of "Yes" of those coded as "honest" increases.

4.4. What Explains the Lacking Effect of the Individualized Component?

The finding that Nudge MDI is not more effective than Nudge MD contradicts our expectations. There are two potential factors that may reduce the effectiveness of Nudge MDI compared to Nudge MD. First, some individuals receiving Nudge MDI in R6 may feel that they have already signaled their dishonesty and that it is too late to rectify their reputation. Second, while Nudge MD is not tailored to the individual, it does contain an element of individual targeting. To dig deeper into where the treatment effects are coming from, we plot the number of "Yes" responses in period 1 against the number of "Yes" responses in period 2 in Figure A.5. A nudge in R1 lowers the fraction that reports a high – in particular, the highest possible – outcome in both periods. We observe a shift from the top right (reporting high outcomes in P1 and P2) to the bottom left (reporting low outcomes in P1 and P2). In contrast, a nudge in R6 induces some of those reporting a high outcome in P1 to report a lower outcome in P2, i.e., a shift from top to bottom right. The evidence does not support the idea that individuals who cheated in P1 feel that it is too late to rectify their reputation but rather suggests that Nudge MD is already perceived as having an individual component.

4.5. Regression Results and the Effect of Guilt/Shame

Finally, we examine if our results hold when we control for different sets of variables in the regressions. Our dependent variable is the number of times subjects report that they guessed correctly in each of the two periods (five rounds). We run pooled ordinary least squares (OLS) regressions.[15] In Table 2, Column (1) we control for period, treatments, and the interaction of period and treatments. In Column (2), we also control for socio-demographic information. In Column (3), we further add controls on self-reported risk tolerance, shame, and guilt.

In line with earlier results, we find that Nudge M in R1 and Nudge MD in R1 decrease the number of reported "Yes" in the first period (coefficients "Nudge M R1" and "Nudge MD

[14] This analysis also shows that a longer text is related to longer time spend on the page, suggesting that subjects actually read the texts.

[15] The results are unchanged when we run ordered probit models or random-effects ordered probit models, to consider the panel structure of our data. We choose to report OLS to ease the interpretation of results.

Table 2: Regression Results

	(1)	(2)	(3)
	b/se	b/se	b/se
Period 2 (R1-R5)	0.157	0.157	0.157
	(0.128)	(0.124)	(0.119)
Nudge M R1	-0.327*	-0.348**	-0.316**
	(0.128)	(0.125)	(0.120)
Nudge MD R1	-0.663***	-0.632***	-0.622***
	(0.128)	(0.125)	(0.120)
Nudge M R6	0.072	0.095	0.099
	(0.128)	(0.125)	(0.120)
Nudge MD R6	-0.085	-0.039	0.001
	(0.128)	(0.125)	(0.120)
Nudge MDI R6	-0.170	-0.144	-0.120
	(0.128)	(0.125)	(0.120)
Period 2 * Nudge M R1	-0.078	-0.078	-0.078
	(0.181)	(0.176)	(0.169)
Period 2 * Nudge MD R1	-0.178	-0.178	-0.178
	(0.182)	(0.177)	(0.169)
Period 2 * Nudge M R6	-0.628***	-0.628***	-0.628***
	(0.181)	(0.176)	(0.169)
Period 2 * Nudge MD R6	-0.716***	-0.716***	-0.716***
	(0.181)	(0.176)	(0.169)
Period 2 * Nudge MDI R6	-0.753***	-0.753***	-0.753***
	(0.181)	(0.176)	(0.169)
Age in yrs.		-0.021***	-0.014***
		(0.002)	(0.002)
Female		-0.176***	-0.014
		(0.052)	(0.051)
Experience online work in h		0.011***	0.009***
		(0.002)	(0.002)
Risk tolerance			0.125***
			(0.010)
Shame			-0.135***
			(0.022)
Guilt			-0.080***
			(0.018)
Additional controls	NO	YES	YES
R2	0.038	0.092	0.167
N	3488	3488	3488

Note: Pooled OLS regressions. The dependent variable is the number of times an individual reports guessing correctly in a set of five rounds. Experience online is a continuous variable measuring the number of hours an individual reports working on online platforms. Risk tolerance is reported on an 11-point scale, where 0 means 'not at all willing to take risks.' Guilt and shame are reported on a 7-point scale, where 1 means that individuals report it is very unlikely that they would experience shame/guilt in the described situation. Additional controls are for income (ordinal), no income information available, or experienced a loss of job/income due to the COVID-19 pandemic.

R1"). The effect of Nudge MD R1 is significant at a higer level and twice as large. The non-significant coefficients on "Nudge M R6", "Nudge MD R6", and "Nudge MDI R6" show that behavior in the first period is the same as in the baseline treatment if nudges are implemented in the middle of the game. The interaction of the R6 nudges with Period 2 shows that the number of "Yes" responses decreases significantly and substantially in the last five rounds. Confirming earlier results, the coefficients on the interaction with Nudge MD R6 and Nudge MDI R6 are higher than with Nudge M R6.

Older individuals are less likely to lie. Those that spend more time earning money online claim a higher number of correct guesses, suggesting that experience increases dishonesty. Self-reporting to be more risk tolerant is positively correlated to the number of Yes. Suffering to a larger extent from shame and guilt is negatively correlated with dishonesty.[16]

Since our results suggest that the effect of nudges is largely independent of the timing, we use the regressions to examine if it is preferable to implement them early. This would be the case if their effect does not deteriorate over time. The coefficients "Period 2 * Nudge M R1" and "Period 2 * Nudge MD R1" in Table 2 show that with an early nudge, second period behavior is not significantly different from first period behavior, suggesting that in our setting it is preferable to implement nudges early. Figure 3 also shows that the effect of the R1 nudges is persistent across periods.[17] This is also confirmed at the round level in Figure A.1.

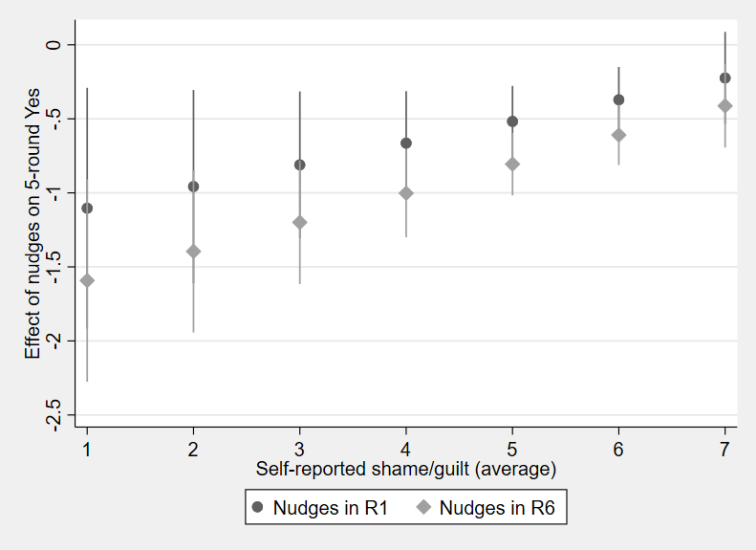
Finally, to provide suggestive evidence on the mechanisms for why our nudges are effective, we examine if the reaction to the nudges varies with self-reported feelings of guilt and shame. Since the two variables are highly correlated (r=0.5), we use their average. We run regressions on the number of reported correct guesses in the first five rounds after nudging, in which we interact being nudged with guilt/shame. For the sake of power, we collapse the content dimension of the nudges. The marginal effects reported in Figure 8 show that suffering to a larger extent from guilt or shame (higher values) is related to lower responsiveness to nudges, suggesting that those who face higher costs of lying do not need a nudge.[18] This might be taken as evidence that the honest types incur an infinite cost of lying while the dishonest types face a finite cost, as discussed in Kajackaite and Gneezy (2017).

[16] According to the psychological literature, there are two schools of thought regarding the difference between guilt and shame. The first school posits that with guilt, the focus is on one's behavior, whereas, with shame, the focus is on one's self. The second school posits that failures that are not publically exposed elicit feelings of guilt, whereas failures that are publically exposed elicit feelings of shame (Cohen et al., 2011).

[17] Nudge M in R1 decreases the number of "Yes" responses in the last five periods from 2.59 in the baseline to 2.19, the difference of 0.39 is significant (p: 0.004). Nudge MD in R1 decreases the number of "Yes" responses in period 2 to 1.75, the difference of 0.84 is again significant (p: 0.000). A comparison of "Yes" responses across all ten rounds shows that it does not make a difference if Nudge M is implemented in R1 or R6 (average number of "Yes" with Nudge M R1: 4.3; Nudge M R6: 4.5; p: 0.137). It may yet be preferable to implement Nudge MD in R1 (average number of "Yes" with Nudge MD R1: 3.5; Nudge MD R6: 4.1; p: 0.002).

[18] We examine if participants' perception to what extent they feel observed varies across treatments and do not find a difference. It seems that in our setting, the degree of observation is unaffected by nudges.

Figure 8: Effect of Nudges Varies with Guilt/Shame



Note: Effects of a nudge in R1 or R6 (collapsing the content dimension) is conditional upon self-reported guilt/shame (average of the two variables), with 1 implying very low and 7 very high feelings of guilt/shame. The dots represent the marginal effects, and the lines stand for the 95% confidence intervals, obtained from OLS regressions in which the dependent variable is the reported number of correct guesses in the five rounds after the nudge (i.e., separate regressions for R1 and R6 nudges). The explanatory variables are a treatment dummy (equal to one if any nudge was provided), self-reported guilt/shame, the interaction thereof, and additional controls as in column (2) of Table 2.

5. Conclusion and Discussion

The question of whether and when honesty nudges are an effective tool for decreasing dishonesty has attracted increasing attention in the literature. In many settings, individuals are nudged after they have already made decisions, and possibly gave in to the temptation to cheat. Examples are taxpayers that submit the annual tax declaration or researchers that have to provide ethics declarations during submission processes. We are the first to examine if nudges work differently when they are implemented before or after decisions have been made.

In line with the idea that nudges cause liars to engage in moral cleansing or conscience accounting, we find that those who were dishonest before nudging are more likely to decrease dishonesty after reading our nudges. While we find that the effect of nudging is quantitatively higher in R6 than in R1, the differences are at most significant at the 10% level. We conclude that the timing of nudges does not matter for their success. Thus, our findings are in line with the recent study by Kristal et al. (2020) showing that it does not matter if an honesty statement is presented at the beginning or at the end of a self-report form. Our results suggest that in settings in which the effect of nudge interventions does not deteriorate over time, as in ours, they should be implemented as early as possible.

We find that the effectiveness of nudges depends on their content. Stressing the possibility of monitoring further reduces dishonesty. In line with the results of field experiments on tax evasion (e.g., Antinyan and Asatryan, 2019), deterrence significantly increases the effectiveness of nudges. Our results show that deterrence can be effective even if there is no formal sanction involved, or when fraud is only suspected and not proven with certainty. However, adding individualized information has no significant additional effect, which is in line with Apesteguia et al. (2013) who show that library users are not more responsive when they are reminded that they were late returning items. There are two potential factors that may reduce the effectiveness of Nudge MDI compared to Nudge MD. First, some individuals may have felt that they already signaled their dishonesty, and that it was too late to rectify their reputation. Second, while Nudge MD is not tailored to the individual, it does contain an element of individual targeting. We find no evidence supporting the idea that individuals who cheated before being nudged felt that it is too late to rectify their reputation. The results suggest that Nudge MD is already perceived as having an individual component.

Nudges are often the only possible intervention when cheating is not observable or monitoring is very costly. We show that in such settings, highlighting the possibility of comparing reported outcomes to chance can be an effective means to reduce cheating. Information about the possibility to compare to chance may be useful, for example, for preventing students from cheating on exams (especially those taking place online), for dissuading researchers from strategically adjusting results, or for deterring taxpayers from misreporting earnings or deductions. In these settings, it is possible to infer if behavior deviates from chance. Examining the effectiveness of nudges exploiting possibilities to compare to chance in real-world settings is an interesting direction for future research.

The magnitude of the effects of our nudges are larger than what other comparable studies find (e.g., Heinicke et al., 2019; Jacquemet et al., 2021). A possible explanation is that our nudges were implemented just before, or even during, the game, while most previous studies

implement them before the experiment starts. Relatedly, Mulder et al. (2020) find that specific rules are more successful for decreasing unethical behavior than general rules. Importantly, we are mainly interested in comparing the effect of nudges implemented before or during the game, so that possible objections to the implementation of nudges during the experiment cancel out. If this result is not driven by other differences in the design of our and other studies, it suggests that a nudge should be implemented in close relation with the task.

The mechanisms discussed in this paper mirror those that are analyzed in the public-goods game literature. In the public goods game as well as in our cheating game, there is a trade-off between self-interest and the moral action. Similar to what has been found in cheating games, free-riding on the provision of public goods has been shown to be conditional. Cooperation is encouraged when participants are allowed to monitor each other and punish non-cooperative behavior (e.g., Fehr and Gächter, 2000; Cinyabuguma et al., 2005; Carpenter et al., 2006; Bochet et al., 2006). Similarly, suasion and nudges have been shown to affect contributions (e.g., Reeson and Tisdell, 2008; Dal Bó and Dal Bó, 2014; Barron and Nurminen, 2020). Dal Bó and Dal Bó (2014) find that punishments and moral messages interact to sustain cooperation. It is possible that the higher effectiveness of Nudges MD and MDI is caused by the interaction of these factors, as they contain both a moral and a deterrence component. Yet, the public goods game and our cheating game are also distinct in important dimensions. Most importantly, in our game, payoffs are not affected by others' decisions, while this is the case in public goods games. Whether moral suasion or deterrence is more or less effective when payoffs are interrelated could be examined in future research.

A possible objection to our study is that, in line with an expanding body of literature (see the recent review by Abeler et al., 2019), our game is simple, does not reflect real life, and, therefore, lacks external validity. However, Potters and Stoop (2016) and Dai et al. (2017) show that simple games like ours predict behavior in the field.

References

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for truth-telling. Econometrica 87(4), 1115–1153.

Antinyan, A. and Z. Asatryan (2019). Nudging for tax compliance: A meta-analysis. ZEW-Centre for European Economic Research Discussion Paper (19-055).

Apesteguia, J., P. Funk, and N. Iriberri (2013). Promoting rule compliance in daily-life: Evidence from a randomized field experiment in the public libraries of barcelona. European Economic Review 64, 266–284.

Barfort, S., N. A. Harmon, F. Hjorth, and A. L. Olsen (2019). Sustaining honesty in public service: The role of selection. American Economic Journal: Economic Policy 11(4), 96–123.

Barron, K. and T. Nurminen (2020). Nudging cooperation in public goods provision. Journal of Behavioral and Experimental Economics 88, 101542.

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. Journal of Political Economy 76(2), 169–217.

Belot, M., S. Choi, J. C. Jamison, N. W. Papageorge, E. Tripodi, and E. Van den Broek-Altenburg (2020). Six-country survey on covid-19. IZA Discussion Paper No. 13230.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating online labor markets for experimental research: Amazon. com's mechanical turk. Political Analysis 20(3), 351–368.

Berinsky, A. J., M. F. Margolis, and M. W. Sances (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. American Journal of Political Science 58(3), 739–753.

Bilen, E. and A. Matros (2021). Online cheating amid covid-19. Journal of Economic Behavior & Organization 182, 196–211.

Bochet, O., T. Page, and L. Putterman (2006). Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior and Organization 60(1), 11–26.

Bott, K. M., A. W. Cappelen, E. Ø. Sørensen, and B. Tungodden (2020). You've got mail: A randomized field experiment on tax evasion. Management Science 66(7), 2801–2819.

Bryan, C. J., G. S. Adams, and B. Monin (2013). When cheating would make you a cheater: implicating the self prevents unethical behavior. Journal of Experimental Psychology: General 142(4), 1001.

Bursztyn, L., S. Fiorin, D. Gottlieb, and M. Kanz (2019). Moral incentives in credit card debt repayment: Evidence from a field experiment. Journal of Political Economy 127(4), 1641–1683.

Cagala, T., U. Glogowsky, and J. Rincke (2021). Detecting and preventing cheating in exams: Evidence from a field experiment. Journal of Human Resources, 0620–10947R1.

Carpenter, J., S. Bowles, and H. Gintis (2006, 05). Mutual monitoring in teams: Theory and experimental evidence on the importance of reciprocity. SSRN Electronic Journal.

Chmielewski, M. and S. C. Kucker (2020). An MTurk Crisis? Shifts in Data Quality and the Impact on Study Results. Social Psychological and Personality Science 11(4), 464–473.

Cinyabuguma, M., T. Page, and L. Putterman (2005). Cooperation under the threat of expulsion in a public goods experiment. Journal of Public Economics 89(8), 1421–1435. The Experimental Approaches to Public Economics.

Cleophas, C., C. Hoennige, F. Meisel, and P. Meyer (2021). Who's cheating? Mining patterns of collusion from text and events in online exams. INFORMS Transactions on Education, ahead of press.

Cohen, T. R., S. T. Wolf, A. T. Panter, and C. A. Insko (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. Journal of personality and social psychology 100(5), 947.

Dai, Z., F. Galeotti, and M. C. Villeval (2017). Cheating in the lab predicts fraud in the field: An experiment in public transportation. Management Science 64(3), 1081–1100.

Dal Bó, E. and P. Dal Bó (2014). "Do the right thing:" The effects of moral suasion on cooperation. Journal of Public Economics 117, 28–38.

DellaVigna, S. and D. Pope (2018). What motivates effort? Evidence and expert forecasts. The Review of Economic Studies 85(2), 1029–1069.

Dimant, E., G. A. Van Kleef, and S. Shalvi (2020). Requiem for a nudge: Framing effects in nudging honesty. Journal of Economic Behavior & Organization 172, 247–266.

Dufwenberg, M. and M. A. Dufwenberg (2018). Lies in disguise - a theoretical analysis of cheating. Journal of Economic Theory 175, 248–264.

Dunaiev, Y. and M. Khadjavi (2021). Collective Honesty? Experimental Evidence on the Effectiveness of Honesty Nudging for Teams. Frontiers in Psychology 12, 2788.

Dwenger, N., H. Kleven, I. Rasul, and J. Rincke (2016). Extrinsic and intrinsic motivations for tax compliance: Evidence from a field experiment in Germany. American Economic Journal: Economic Policy 8(3), 203–32.

Exley, C. L. and J. B. Kessler (2022). The gender gap in self-promotion. The Quarterly Journal of Economics 137(3), 1345–1381.

Fehr, E. and S. Gächter (2000). Cooperation and punishment in public goods experiments. The American Economic Review 90(4), 980–994.

Fellner, G., R. Sausgruber, and C. Traxler (2013). Testing enforcement strategies in the field: Threat, moral appeal and social information. Journal of the European Economic Association 11(3), 634–660.

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in disguise - an experimental study on cheating. Journal of the European Economic Association 11, 525–547.

Galeotti, F., C. Saucet, and M. C. Villeval (2020). Unethical amnesia responds more to instrumental than to hedonic motives. Proceedings of the National Academy of Sciences 117(41), 25423–25428.

Garbarino, E., R. Slonim, and M. C. Villeval (2019). Loss aversion and lying behavior. Journal of Economic Behavior & Organization 158, 379–393.

Gerlach, P., K. Teodorescu, and R. Hertwig (2019). The truth about lies: A meta-analysis on dishonest behavior. Psychological Bulletin 145(1), 1.

Gneezy, U., A. Imas, and K. Madarász (2014). Conscience accounting: Emotion dynamics and social behavior. Management Science 60(11), 2645–2658.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying aversion and the size of the lie. American Economic Review 108(2), 419–53.

Grossman, Z. and J. J. van der Weele (2016, 12). Self-Image and Willful Ignorance in Social Decisions. Journal of the European Economic Association 15(1), 173–217.

Gsottbauer, E., D. Müller, S. Müler, S. T. Trautmann, and G. Zudenkova (2022). Social class and (un)ethical behaviour: Causal and correlational evidence. The Economic Journal, ahead of print.

Heinicke, F., S. Rosenkranz, and U. Weitzel (2019). The effect of pledges on the distribution of lying behavior: An online experiment. Journal of Economic Psychology 73, 136–151.

Holz, J. E., J. A. List, A. Zentner, M. Cardoza, and J. Zentner (2020). The $100 million nudge: Increasing tax compliance of businesses and the self-employed using a natural field experiment. Technical report, National Bureau of Economic Research.

Hüllemann, S., G. Schüpfer, and J. Mauch (2017). Application of Benford's law: A valuable tool for detecting scientific papers with fabricated data? Der Anaesthesist 66(10), 795–802.

Ilies, R., A. C. Peng, K. Savani, and N. Dimotakis (2013). Guilty and helpful: An emotion-based reparatory model of voluntary work behavior. Journal of Applied Psychology 98(6), 1051.

Jacquemet, N., A. G. James, S. Luchini, J. J. Murphy, and J. F. Shogren (2021). Do truth-telling oaths improve honesty in crowd-working? PloS one 16(1), e0244958.

Jacquemet, N., S. Luchini, J. Rosaz, and J. F. Shogren (2019). Truth telling under oath. Management Science 65(1), 426–438.

Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. Journal of Economic Behavior & Organization 93, 328–336.

Kajackaite, A. and U. Gneezy (2017). Incentives and cheating. Games and Economic Behavior 102, 433–444.

Kennedy, R., S. Clifford, T. Burleigh, P. D. Waggoner, R. Jewell, and N. J. G. Winter (2020). The shape of and solutions to the mturk quality crisis. Political Science Research and Methods 8(4), 614–629.

Khalmetski, K. and D. Sliwka (2019, November). Disguising Lies - Image Concerns and Partial Lying in Cheating Games. American Economic Journal: Microeconomics 11(4), 79–110.

Kiesler, C. A. (1971). The psychology of commitment: Experiments linking behavior to belief. Academic Press Inc.

Kristal, A. S., A. V. Whillans, M. H. Bazerman, F. Gino, L. L. Shu, N. Mazar, and D. Ariely (2020). Signing at the beginning versus at the end does not decrease dishonesty. Proceedings of the National Academy of Sciences 117(13), 7103–7107.

Lin, M.-J. and S. D. Levitt (2020). Catching cheating students. Economica 87(348), 885–900.

Mazar, N., O. Amir, and D. Ariely (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. Journal of Marketing Research 45, 633–644.

Mills, S. (2020). Personalized nudging. Behavioural Public Policy, 1–10.

Mulder, L. B., F. Rink, and J. Jordan (2020). Constraining temptation: How specific and general rules mitigate the effect of personal gain on unethical behavior. Journal of Economic Psychology 76, 102242.

Ploner, M. and T. Regner (2013). Self-image and moral balancing: An experimental analysis. Journal of Economic Behavior & Organization 93, 374–383.

Potters, J. and J. Stoop (2016). Do cheaters in the lab also cheat in the field? European Economic Review 87, 26–33.

Reeson, A. F. and J. G. Tisdell (2008). Institutions, motivations and public goods: An experimental test of motivational crowding. Journal of Economic Behavior & Organization 68(1), 273–281.

Rotella, A., J. Jung, C. Chinn, and P. Barclay (2019). Observation and moral ambiguity matter: A meta-analysis on moral licensing. Technical report.

Savić, M., J. Atanasijević, D. Jakovetić, and N. Krejić (2022). Tax evasion risk management using a hybrid unsupervised outlier detection method. Expert Systems with Applications 193, 116409.

Shalvi, S. and C. K. De Dreu (2014). Oxytocin promotes group-serving dishonesty. Proceedings of the National Academy of Sciences 111(15), 5503–5507.

Shalvi, S., F. Gino, R. Barkan, and S. Ayal (2015). Self-serving justifications: Doing wrong and feeling moral. Current Directions in Psychological Science 24(2), 125–130.

Shu, L. L., F. Gino, and M. H. Bazerman (2011). Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting. Personality and social psychology bulletin 37(3), 330–349.

Tödter, K.-H. (2009). Benford's law as an indicator of fraud in economics. German Economic Review 10(3), 339–351.

Verschuere, B., E. H. Meijer, A. Jim, K. Hoogesteyn, R. Orthey, R. J. McCarthy, J. J. Skowronski, O. A. Acar, B. Aczel, B. E. Bakos, et al. (2018). Registered replication report on Mazar, Amir, and Ariely (2008). Advances in Methods and Practices in Psychological Science 1(3), 299–317.

West, C. and C.-B. Zhong (2015). Moral cleansing. Current Opinion in Psychology 6, 221–225. Morality and ethics.

Zhao, J., Z. Dong, and R. Yu (2019). Don't remind me: When explicit and implicit moral reminders enhance dishonesty. Journal of Experimental Social Psychology 85, 103895.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. Experimental Economics 13(1), 75–98.

Appendix

A.1. Questionnaire

Welcome Screen:
Welcome!
You will now take part in a research study from Freiburg and Rennes University. With your participation you will make an important contribution to research. The survey is about individuals' memory and behavior in situations marked by randomness. You will participate in two tasks and answer a brief survey. Participation should take about 10 minutes. It is very important that you read the instructions carefully and that you complete the entire study, once you have started. For completing the survey, you will receive a fixed payment of $1. In the second task, you have the chance to earn up to $2 in addition to your fixed payment depending on your choices and on the results of random draws. Additional payments will be distributed as a bonus payment within three days upon completion of the study.
Please note that you will have to answer control questions correctly to get the completion code. If you fail any of them, the survey will automatically end. Confidentiality: You will remain anonymous throughout the study. Results may include summary data, but you will never be identified. All data collected are for research purposes only. By continuing, you consent to the publication of study results. Participants' rights: Your participation is voluntary. You may stop participating at any time by closing the browser window or the program to withdraw from the study. If you stop participating in this survey before completing it, please, kindly contact us at [...] if you do not want your data to be used.
Please also contact us at [...] if you have any questions. By continuing, you indicate that you are at least 18 years old, have read and understood the above, and you voluntarily agree to participate in this study.

Screener Question:
Before we proceed, we have a question about how you are feeling.
Recent research on decision making shows that choices are affected by the situational context. Differences in how people feel and their environment can affect choices. To help us understand how people make decisions, we are interested in information about you. Specifically, we would like to know if you actually take time to read the instructions; if not, some results may not tell us very much about decision making in the real world. To show that you have read the instructions, please ignore the question below how you are feeling and instead check only the 'none of the above' option as your answer.
Please check all words that describe how you are currently feeling
Interested, Distressed, Excited, Upset, Strong, Guilty, Scared, Hostile, Enthusiastic, Proud, Irritable, Alert, Ashamed, Inspired, Nervous, Determined, Attentive, Jittery, Active, Afraid, None of the above

Memory Game:
You will now participate in three rounds of a memory game. You will be shown sequences of

letters of increasing lengths, see the example below.
Your task is to remember the sequence and enter it on the next screen. Each of the three rounds proceeds as follows:

- You are shown a sequence of letters for 10 seconds and are asked to remember it.
- You enter as many letters as you can remember.

Please note that you are able to skip the timer if you are ready to enter the letters.
Are you ready to start the game? If yes, please continue.

page break

Here is the sequence that you need to remember (repeated for 3 rounds)

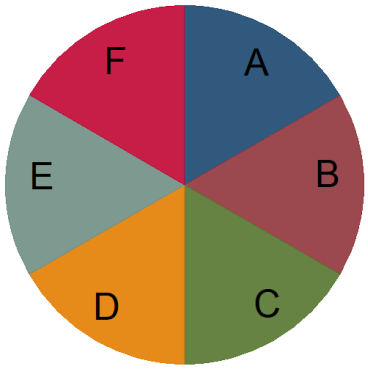
page break

Please enter the letters that you can remember below

Are you ready to start the next round? If yes, please continue.

page break

Cheating Game:
Task 2: You can now earn a bonus
You will now participate in 10 rounds of a wheel of fortune. In each round you earn either $0 or $0.20. The wheel of fortune shows the first 6 letters of the alphabet, from A to F:

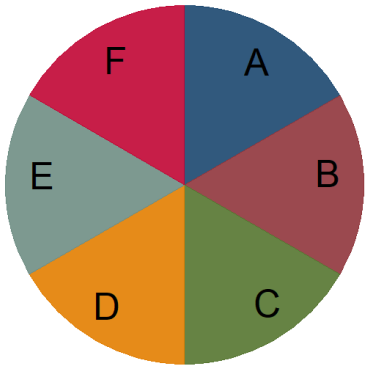


Your task is to guess which letter the wheel of fortune will show and report if your guess was correct. Each of the 10 rounds proceeds as follows: You have 10 seconds to guess which letter the wheel will show and remember it. The virtual wheel is started and one of the letters (from A to F) is randomly selected. You report if the letter you guessed is equal to the letter selected by the wheel If the guessed and the selected letter are the same: you report “Yes” and receive $0.20. If the guessed and the selected letter are NOT the same: you report “No” and receive $0. Please note that you are able to skip the timer if you are ready to spin the wheel.

Before you continue, we would like to ascertain that you understand the rules. You have to answer the two following questions correctly in order to get the completion code.
What is your payment if the letter that you guessed is the same as the selected letter?
$0.0, $0.1, $0.2, $0.3
What is your payment if the letter that you guessed is NOT the same as the selected letter?
$0.0, $0.1, $0.2, $0.3

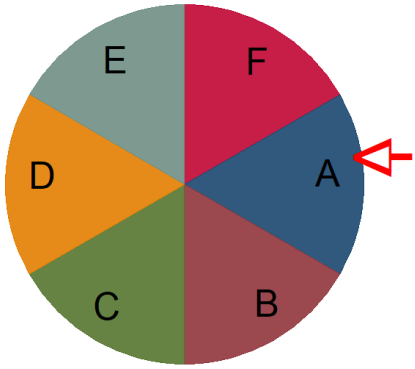
page break

[repeated for ten rounds]
Now you have 10 seconds to guess which letter the wheel will show and remember it.
Have you made your guess?



page break

[Example]
Here is the result of your wheel spin (the red arrow indicates the selected letter):



Is the letter that you guessed in this round the same as the selected letter? Yes/No

Follow-Up Survey:

- What is your gender? Male, Female, Other, Prefer not to say
- What is your year of birth?
- What is the highest level of education you have completed? Less than high school degree, High school graduate (high school diploma or equivalent including GED), Some college but no degree, Associate degree in college (2-year), Bachelor's degree in college (4-year), Master's degree, Doctoral degree, Professional degree (JD, MD), Prefer not to say
- What is your current employment status? Full-time employee, Part-time employee, Self-employed or business owner, Unemployed and looking for work, Student, Not in labor force (e.g., retired, full-time parent), Other, Prefer not to say
- In what range was your total household income in 2019 before taxes? Less than $10,000, $10,000 to $19,999, $20,000 to $29,999, $30,000 to $39,999, $40,000 to $49,999, $50,000 to $59,999, $60,000 to $69,999, $70,000 to $79,999, $80,000 to $89,999, $90,000 to $99,999, $100,000 to $149,999, $150,000 or more, Prefer not to say
- How many hours per week do you spend on platforms such as Mturk doing tasks for money?
- Have you lost your job or has your activity (as self-employed) been stopped as a consequence of the Covid-19 pandemic? Yes, No
- Have you experienced a fall in houshold income as a consequence of the Covid-19 pandemic? Yes, No
- Have you experienced any non-financial effects from the societal changes occuring due to the Covid-19 pandemic, such as (select all that apply): Enjoying more free time, Enjoying time with family, Reduction of air pollution, Reduction of noise pollution, Boredom, Loneliness, Increased conflicts with relatives, friends, neighbours, General anxiety and stress
- How do you see yourself: are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please tick a box on the scale, where the value 0 means: 'not at all willing to take risks' and the value 10 means: 'very willing to take risks'.
- Assume that you make a mistake and find out someone else is blamed for the error. Later, this person confronts you about your mistake. Please indicate on a scale where 1 means 'very unlikely' and 7 means 'very likely', how likely it is that you would feel terrible in this situation?
- Assume that you lie to people but they never find out about it. Please indicate on a scale where 1 means 'very unlikely' and 7 means 'very likely', how likely it is that you would feel terrible about the lies told?

- To what extent did you feel observed during the study? To a large extent, To a moderate extent, To a small extent, Not at all, Prefer not to say
- What do you think we are trying to assess with this study?

A.2. Additional Results

Table A.1: Sample Characteristics by Treatment

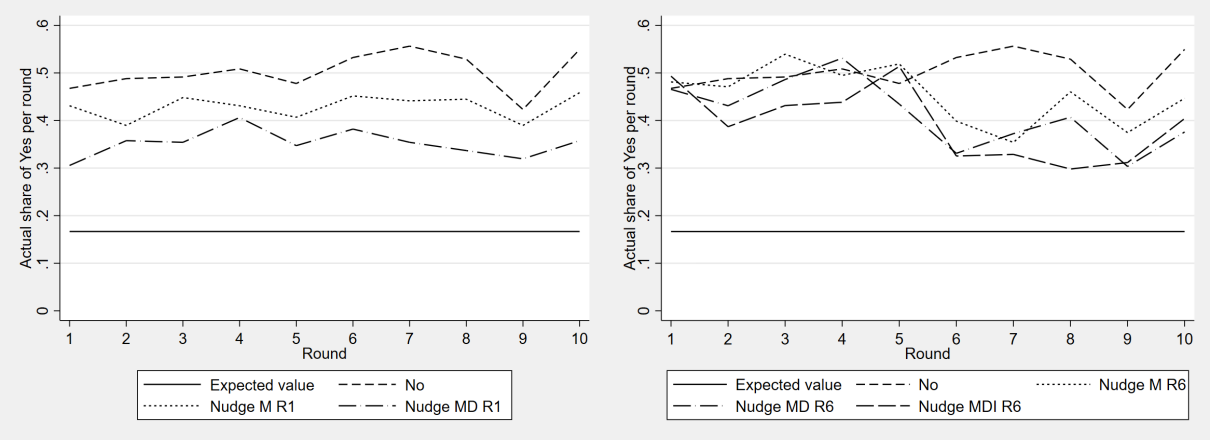
	Nudge/Treatment						
	No	M R1	MD R1	M R6	MD R6	MDI R6	Total
Age	38.28	37.84	38.74	38.14	39.99	38.22	38.53
	(10.50)	(10.83)	(11.99)	(11.59)	(12.40)	(11.25)	(11.45)
Female	0.45	0.40	0.42	0.45	0.41	0.45	0.43
	(0.50)	(0.49)	(0.49)	(0.50)	(0.49)	(0.50)	(0.50)
Work online in h	20.78	19.90	19.97	20.10	19.41	18.79	19.83
	(17.43)	(14.62)	(14.40)	(14.18)	(13.50)	(14.33)	(14.79)
N	293	290	288	291	290	292	1744

Note: The means are reported with the standard deviations in parentheses.

Figure A.1: Actual Share of "Yes" Responses by Round and Treatment

(a) No and R1 treatments

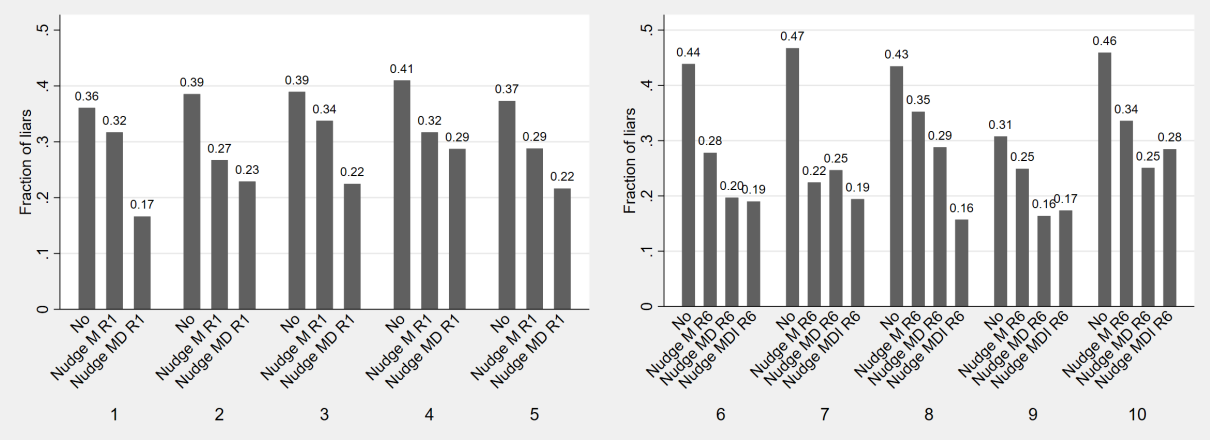
(b) No and R6 treatments



Note: The expected number of "Yes" responses is 0.167, and indicated by the horizontal line. Nudge M is a moral reminder, Nudge MD adds a deterrent component to Nudge M, Nudge MDI adds individualized information to Nudge MD. The numbers of obs per round are as follows: No: 293, Nudge M R1: 290, Nudge MD R1: 288, Nudge M R6: 291, Nudge MD R6: 290, Nudge MDI R6: 292.

Figure A.2: Share of Liars by Round and Treatment

(a) Nudges in R1 (b) Nudges in R6



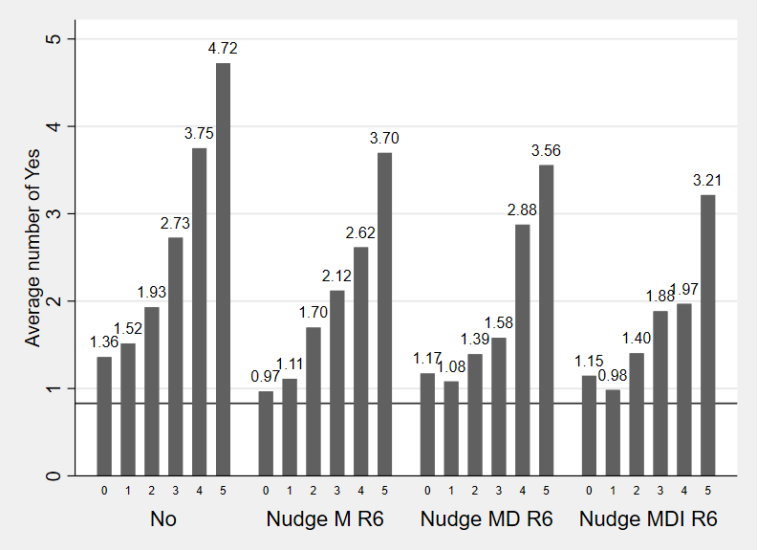
Note: For each round and treatment, we report the share of liars, calculated as: fraction of "Yes" in a round-0.167/(1-0.167). We focus on the first five rounds after nudging.

Table A.2: Time Spent on a Nudge Screen

Nudge	Seconds
Nudge M R1	9.34
Nudge M R6 Dishonest	9.67
Nudge M R6 Honest	8.32
Nudge MD R1	13.83
Nudge MD R6 Dishonest	13.75
Nudge MD R6 Honest	13.58
Nudge MDI R6 Dishonest	19.14
Nudge MDI R6 Honest	18.33

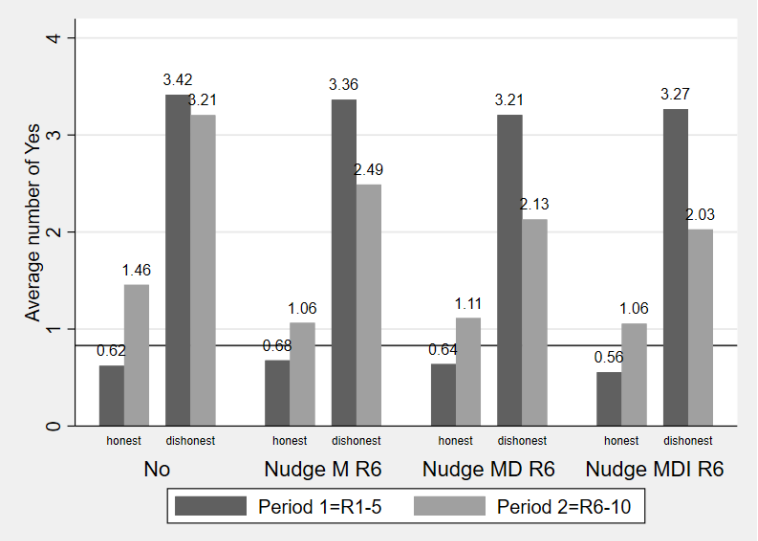
Note: Reported are the seconds spent on the screen with the nudge. (Presumably) honest participants are those who reported 0 or 1 "Yes" in period 1, and dishonest participants are those who reported 2, 3, 4, or 5 "Yes".

Figure A.3: Behavior by Number of Correct Guesses in the First Period



Note: 0, 1, 2, 3, 4 and 5 refer to the number of times an individual reports guessing correctly in the first period (rounds 1 to 5). Reported is the average number of times individuals reported having being correct in the second period. See also the note for Figure 7.

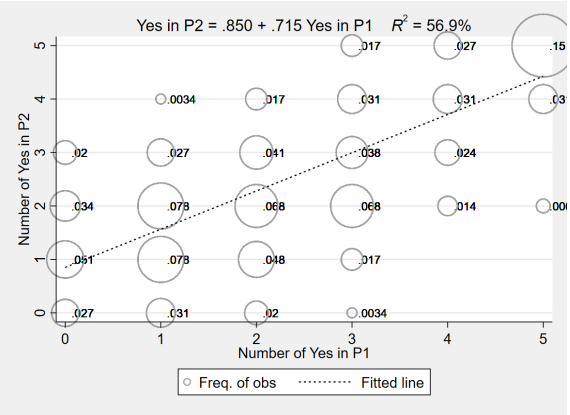
Figure A.4: Robustness of Figure 7: Coding 2 as Dishonest



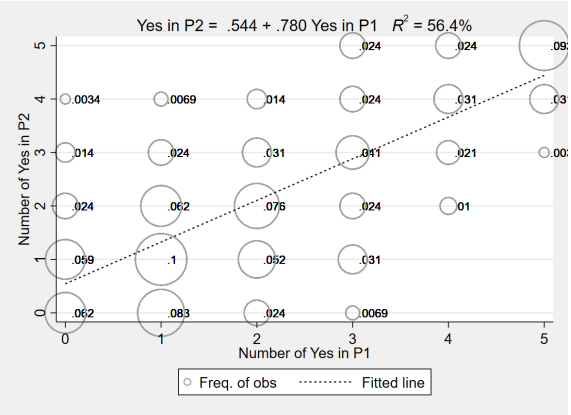
Note: Reported are the number of times an individual reports guessing correctly within five rounds. The expected number of "Yes" is 0.8225, and indicated by the horizontal line. (Presumably) honest participants are those who reported 0 or 1 "Yes" in period 1, and dishonest participants are those who reported 2, 3, 4, or 5 "Yes".

Figure A.5: Correlation Between Periods

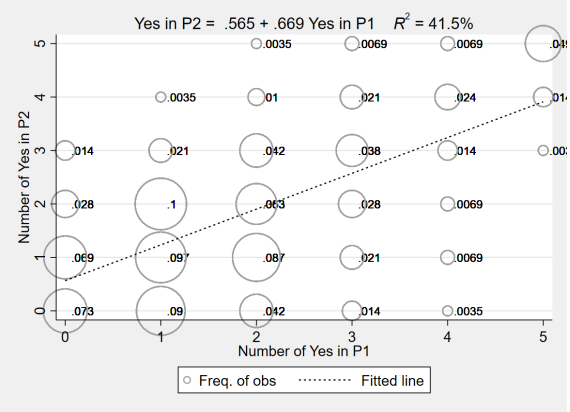
(a) Control treatment



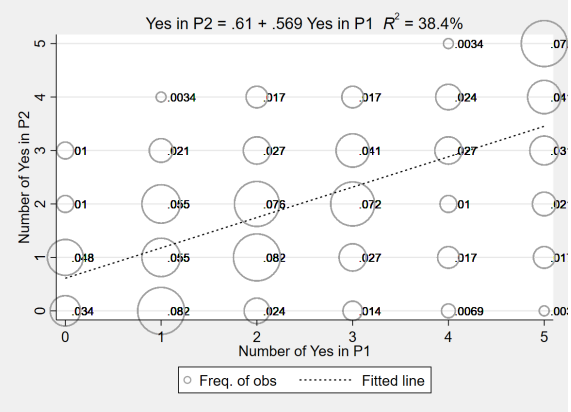
(b) Nudge M R1



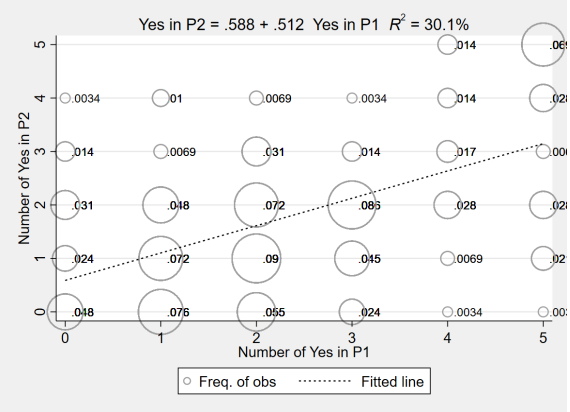
(c) Nudge MD R1



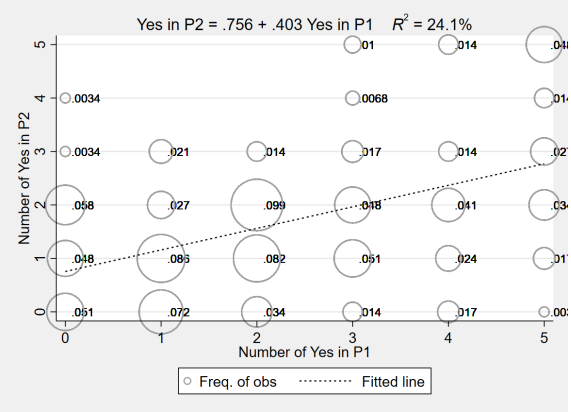
(d) Nudge M R6



(e) Nudge MD R6



(f) Nudge MDI R6



Note: The circles present the frequency of the observations. The numbers indicate the fraction of observations in a cell relative to all subjects in the treatment. Each line presents the line of best fit using simple OLS regression.