Dimant, Eugen; van Kleef, Gerben A.; Shalvi, Shaul

Working Paper

Requiem for a Nudge: Framing Effects in Nudging Honesty

CESifo Working Paper, No. 8170

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

Suggested Citation: Dimant, Eugen; van Kleef, Gerben A.; Shalvi, Shaul (2020) : Requiem for a Nudge: Framing Effects in Nudging Honesty, CESifo Working Paper, No. 8170, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/216566

CESifo WORKING PAPERS

8170
2020

March 2020

Requiem for a Nudge: Framing Effects in Nudging Honesty

Eugen Dimant, Gerben A. van Kleef, Shaul Shalvi

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. 8170*

Requiem for a Nudge: Framing Effects in Nudging Honesty

Abstract

We examine framing effects in nudging honesty, in the spirit of the growing norm-nudge literature, by utilizing a high-powered and pre-registered study. Across four treatments, participants received one random truthful norm-nudge that emphasized 'moral suasion' based on either what other participants previously did (empirical message) or approved of doing (normative message) and varied in the framing (positive or negative) in which it was presented. Subsequently, participants repeatedly played the 'mind game' in which they were first asked to think of a number, then rolled a digital die, and then reported whether the two numbers coincide, in which case a bonus was paid. Hence, whether or not the report was truthful remained unobservable to the experimenters. We find compelling null effects with tight confidence intervals showing that none of the norm-nudge interventions worked. A follow-up experiment reveals the reason for these convincing null-effects: the information norm-nudges did not actually change norms. Notably, our secondary results suggest that a substantial portion of individuals misremembered norm-nudges such that they conveniently supported deviant behavior. This subset of participants indeed displayed significantly higher deviance levels, a behavior pattern in line with literature on motivated misremembering and belief distortion. We discuss the importance of this high-powered null finding for the flourishing norm-nudge literature and derive policy implications.

JEL-Codes: B410, D010, D900.

Keywords: norm-nudges, nudge, social information, social norms.

*Eugen Dimant**
University of Pennsylvania / Philadelphia / USA
edimant@sas.upenn.edu

Gerben A. van Kleef
University of Amsterdam / The Netherlands
G.A.vanKleef@uva.nl

Shaul Shalvi
University of Amsterdam / The Netherlands
S.Shalvi@uva.nl

*corresponding author

This version: February 8, 2020
Forthcoming in Journal of Economic Behavior & Organization. This work benefited from conversations with Johannes Abeler, several researchers at the German Institute for Economic Research (DIW), Daniela Puzzello, and two anonymous reviewers. IRB approval was obtained from the University of Pennsylvania. This work was supported by European Research Council Grant No. ERC-StG-637915. The data collection and analyses were pre-registered at AsPredicted.org #23244 and #23283 entitled "Same same but different? The role of framing in norm-nudge interventions".

1. Introduction

We study the impact of norm-nudges on deviant behavior by creating interventions that are in the spirit of the growing norm-nudge literature (for a recent methodological discussion, see Bicchieri and Dimant, 2019). There, personalized messages – and sometimes physical letters – are being sent with the intent to achieve behavior change in the forms of increasing tax compliance, charitable giving, primary school enrollment, and student learning, or in the form of decreasing energy consumption, student absenteeism from school, credit card debt, and even the spread of HIV infections, among other examples (Reinikka and Svensson, 2005; Dupas, 2011; Luttmer and Singhal, 2014; Boyer et al., 2016; Brandon et al., 2017; Rogers and Feller, 2018; Bursztyn et al., 2019). In existing literature, such information-only interventions indicate mixed results, with some studies demonstrating success (Hallsworth et al., 2017; Hernandez et al., 2017; Bott et al., 2019) while others either fail to detect a significant effect (Blumenthal et al., 2001; Fellner et al., 2013; Castro and Scartascini, 2015; Kettle et al., 2017; Cranor et al., 2018; Dunning et al., 2019) or indicate that interventions may backfire (John and Blume, 2018; Bicchieri et al., 2019c).

For norm-nudging to be effective, one must correctly identify the mechanisms through which different types of information affect behavior. We must understand the specific context in which the targeted behavior occurs (see discussion in Gino et al., 2019). In addition, despite a long tradition in social psychology, philosophy, and economic literature showing when, why, and how individuals react and conform to descriptive and normative information of peers (e.g., Deutsch and Gerard, 1955; Cialdini and Trost, 1998; Gino et al., 2009; Cialdini and Goldstein, 2004; Bicchieri, 2006; Schultz et al., 2007; Bicchieri et al., 2019a; Dimant, 2019; van Kleef et al., 2019), science is still working towards a better understanding of how to properly frame norm-nudges, which is a highly debated topic in ongoing research (see discussions in, e.g., Bicchieri and Dimant, 2019; Larkin et al., 2019). It is important to advance our understanding on how to best frame norm-nudges because better framing can serve as a cost-effective way to increase the effectiveness of nudge interventions. Our high-powered and pre-registered study adds to this literature by separating out the effects of framing and different forms of norm information (descriptive or normative). To the best of our knowledge, our paper is the first to explore this for lying behavior, in particular in the context of the mind game.

A comprehensive literature in economics has advanced our understanding of the mechanisms mediating the extent of deviant behavior, including applicability of social norms, image concerns, preference for appearing honest, and intrinsic lying costs, among others (e.g., Mazar et al., 2008; Hao and Houser, 2017; Gneezy et al., 2018a; Kajackaite, 2018; Bicchieri et al., 2019b). We extend this scholarly debate by studying a particular source of such behavior, namely the impact of norm-nudges on reducing deviant behavior in an environment in which there is no risk of being caught and thus risk perception updating is ruled out by design. To achieve this, we capitalize on the 'cheat in your mind' paradigm (Jiang, 2013) for the reporting decisions in our experiment. Using this paradigm over the congeneric die-paradigm (e.g., Shalvi et al., 2011; Fischbacher and Föllmi-Heusi, 2013) or other related tasks capturing deviant behavior (e.g., Buckenmaier et al., 2019; Dimant, 2019, for an overview see also Abeler et al., 2019; Gerlach et al., 2019; Köbis et al., 2019) has the methodological advantage that participants cannot be worried that their lying behavior is verifiable by anyone, including the experimenter. This is known to matter to the participants, e.g. due to self-/social-image concerns (Bénabou and Tirole, 2006; Andreoni and Bernheim, 2009; Bolton et al., 2019). Relevant to our context, literature has established important differences between paradigms in which lying can and cannot be observed such as responsiveness in cheating behavior to incentives (Kajackaite and Gneezy, 2017).

Across four treatments, participants received one random truthful norm-nudge that emphasized 'moral suasion' based on either what other participants *previously did* (empirical message) or *approved of doing* (normative message). Motivated by existing literature that recognizes the relevance of framing effects (e.g., Andreoni, 1995; Levin et al., 1998; Kahneman, 2003; Alekseev et al., 2017, on the relationship between framing and norms, see Chang et al., 2019), we also varied the framing of the norm messages and presented them in either a positive way (majority did not cheat / did not approve of cheating) or an inverted and equivalent negative way (minority cheated / approved of cheating). Our working assumption is that the *principle of description invariance*, the presumption that logically equivalent descriptions of a situation that differ only in framing (e.g., positive versus negative) will lead to the same choices, does not hold in the realm of social norm interventions designed to nudge individual behavior ("norm-nudge"). Existing research identifies such frame-sensitivity within the medical context (infamously known as the "Asian Disease Problem" Tversky and Kahneman, 1981), altruism (Sonnemans et al., 1998), and conference registration decisions of economists (Gächter et al., 2009). Despite its cost-effective nature, however, a structural understanding of the role of framing is ab-

sent in the nudge literature.[1] In addition, only recently have scholars started evaluating the long-term impacts of nudge interventions (e.g., Brandon et al., 2017, but see also the *Behavior Change for Good Initiative* spearheaded by Angela Duckworth and Katherine Milkman). Because this constitutes an important but unanswered policy question, we add to this debate by measuring the norm-nudge framing in a repeated context in order to study whether and how quickly the impact of norm-nudges dissolves.

We utilize the Amazon Mechanical Turk (MTurk) platform in order to achieve the desired statistical power across all variations of the experiment. Recent literature points to the robustness, generalizability, and reproducibility of laboratory findings in online environments (Arechar et al., 2018; Coppock et al., 2018; Snowberg and Yariv, 2018).[2] Results from our high-powered study suggest a resounding null effect with tight confidence intervals. Our main conclusion is that none of the norm-nudge interventions yielded any significant behavioral differences, with misreporting (= winning reports about expected value of $\frac{1}{6}$) hovering at around 30% in all treatments.

In a follow-up experiment, we capitalize on existing social norms research to examine *why* these information-based norm interventions remained ineffective. By capitalizing on the norm elicitation approach of Krupka and Weber (2013), we test the conjecture that such 'soft' interventions may not be sufficient to actually *shift* an existing norm. This is indeed what we find; the norm-based interventions are unable to shift norm perceptions in the context of our study. This is an important finding in that such norm-nudges are commonly employed, and it can explain why they remain ineffective at times, as discussed above. This finding is also consistent with recent experimental and theoretical evidence suggesting that behavior can be sticky and requires a less gentle 'shove' or an active behavioral intervention (e.g., Kahan, 2000; Grüne-Yanoff and Hertwig, 2016, see also Ambuehl et al., 2019).

[1] In a recent paper, Sunstein (2017, pg. 18) urges caution and emphasizes the relevance of framing in nudge interventions: "Consider an analogy: if a company says that its product is "90% fat-free", people are likely to be drawn to it, far more so than if the company says that its product is "10% fat'.' The two phrases mean the same thing, and the "90% fat-free" frame is legitimately seen as a form of manipulation. In 2011, the American government allowed companies to say that their products are 90% fat-free – but only if they also say that they are 10% fat. We could imagine similar constraints on misleading or manipulative frames that are aimed at getting people to opt out of the default."

[2] MTurk is well suited for our experimental design, which does not involve any direct interaction between participants. Literature related to our research agenda has successfully used variants of cheating paradigms and/or norm-nudge interventions on MTurk (Peer et al., 2014; Hildreth et al., 2016; Bicchieri et al., 2019b; Bolton et al., 2019; Charness et al., 2019). Importantly, a recent meta-study by Gerlach et al. (2019) finds that lying behavior in the dice paradigm does not differ significantly between MTurk and the laboratory.

The paper proceeds as follows: Section 2 details our data collection procedure (2.1), the design (2.2), the testable hypotheses (2.3), and the results (2.4). Section 3 concludes.

2. Experiment

2.1. Data Collection

Our main experiment contains data from a high-powered study[3] that was collected from 1,200 Amazon Mechanical Turk participants in May 2019 containing 52% females and an average age of 38.7 years. The experiment lasted about 9 minutes and participants earned an average of $1.20 (including a show-up payment of $0.30), which translates to an hourly wage of roughly $8 and is well above average on the MTurk platform (Hara et al., 2018).[4]

Our experiment contains one control and four treatments (see Appendix for a breakdown of observations per treatment and Figure 1 for details). A participant took part only in one randomly selected condition, and data for all conditions was collected simultaneously.

2.2. Design

Our design is straightforward and contained two parts: (1) provision of a norm-nudge in which participants observed information pointing towards a social norm of behavior in the experiment, and (2) repeated reporting decisions across 20 periods. In this paradigm as introduced by Jiang (2013), both correctly reporting the winning number and lying behavior -reporting the rolled number despite having thought of a different one- yielded a monetary bonus without imposing any risk of being caught for lying. Multiple comprehension, manipulation, and attention checks throughout the experiment and at the end ensured that our treatment interventions and the observed results were credible.

[3] We calibrate the required sample size to obtain high statistical power based on a pre-test that was run on the campus of UPenn (n=40, effect size 0.387). Ultimately, this allows us to achieve statistical power in excess of 95% – see our pre-registration documents for more details and calculations that can be obtained from https://aspredicted.org/3pi2g.pdf and https://aspredicted.org/7uy3n.pdf. To ensure high quality data collection on mTurk, we utilize a combination of CAPTCHAs and sophisticated screening questions to avoid pool contamination. We applied the following restrictions to the participant pool: participants had to be in the U.S., approval rate greater than 99%, and could participate only once.

[4] To test the robustness of and provide an intuition for our results, we also collected data in two follow-up experiments (n=250 and n=217, respectively) yielding comparable hourly payoffs. We provide details on both experiments in Footnote 9, Figure 4 in combination with discussion in Section 2.4, and the Appendix.

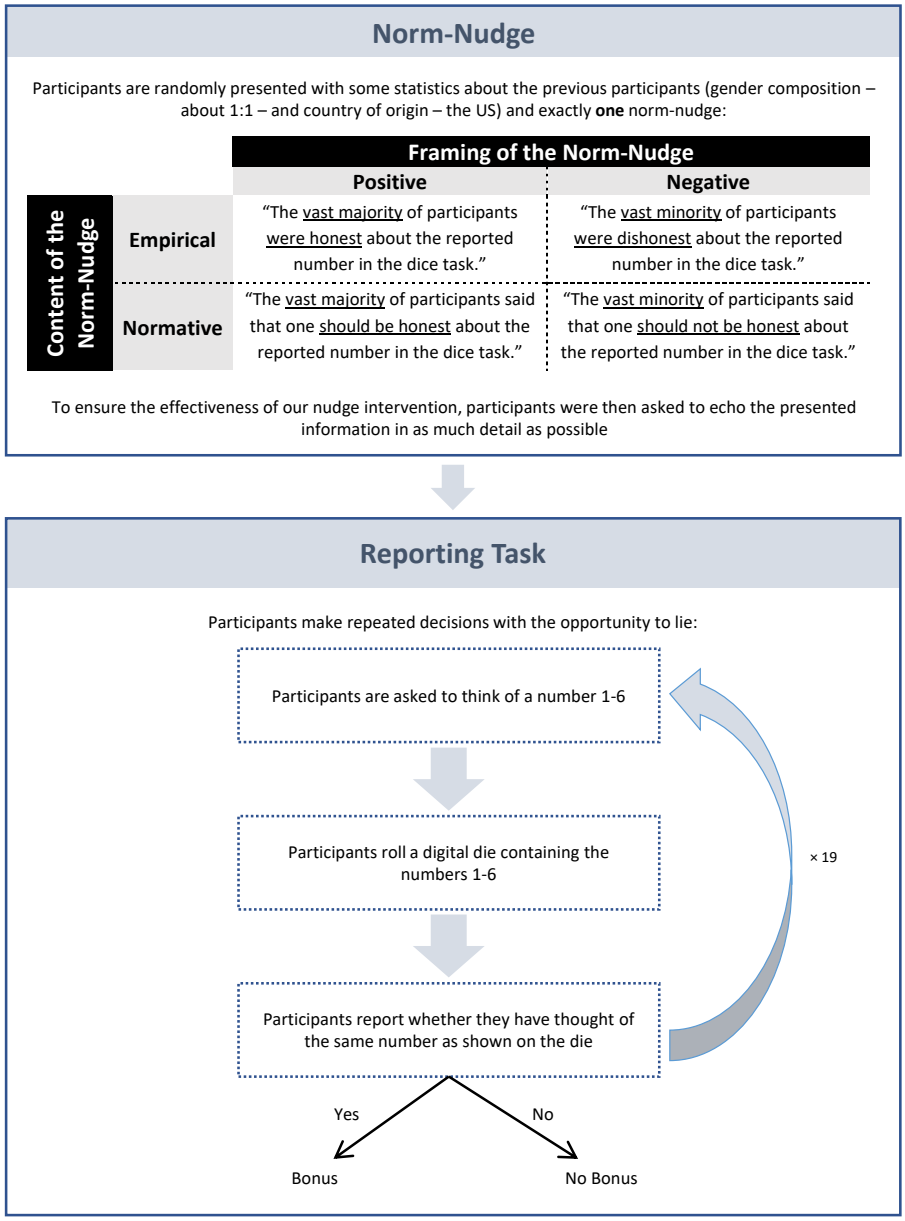

Figure 1: Experimental design. In the control condition, participants received the generic text "previous participants completed the task". After the first iteration, the reporting task was repeated 19 more times.

Part 1: Norm-Nudge

Following the provision of consent, participants were randomly assigned to one of the information treatments and observed exactly three pieces of information. The first two pieces were always the same in all conditions and the third was randomly varied according

to the norm-nudge intervention. The information provided to participants contained: (1) the gender ratio (about 1:1), (2) their location (United States), and (3) one of the four variations of norm information. Random variations in (3) allow us to causally identify treatment differences of the norm-nudge interventions. Participants who were instead randomly allocated to the control condition observed the same information in (1) and (2), but in (3), rather than observing norm information, these participants simply received the note that "previous participants completed the task".

Norm messages varied along two dimensions: content and framing. With respect to the content, we followed the tradition of norm-nudge interventions (see conceptual discussion in Bicchieri and Dimant, 2019). We also followed the social norm tradition of Bicchieri (2006, see also Cialdini and Trost, 1998; Cialdini et al., 2006; Schultz et al., 2007) by distinguishing between *empirical* information (what the majority of other people have done in a similar situation) and *normative* information (what the majority of other people have approved of doing in a similar situation).[5] Our framing variations were inspired by another stream of existing literature (e.g., Tversky and Kahneman, 1981; Levin et al., 1998; Sonnemans et al., 1998; Kahneman, 2003; Gächter et al., 2009; Alekseev et al., 2017) and involved either a positive framing (majority did not cheat / did not approve of cheating) or inverted in a negative way (minority cheated / approved of cheating). As such, framing only affected the way in which the norm-nudge was presented while holding its content constant.

After having acknowledged that all three pieces of information were read carefully, participants proceeded to the next screen where they were asked to repeat the provided information and write it out in as much detail as possible. We used this to increase the salience of the message and as an attention check to make sure that the content of the intervention was correctly observed and retained.[6] Subsequently, participants were *again* displayed the correct norm-nudge information and were asked to acknowledge that they have read and retained it. Importantly, such a minimal exogenous variation of the

[5] To ensure the truthfulness of all norm-nudges, the information was based on the behavior and beliefs of other uninvolved participants from a previous study run at the University of Pennsylvania. Such methodology is commonly adopted in experimental research on social norms (e.g., Goldstein et al., 2008; Bicchieri and Xiao, 2009; Krupka and Weber, 2013; Bicchieri et al., 2019c). The truthful results were that the majority of the participants did not lie and that the majority did not approve of lying.

[6] Particular emphasis was put on whether participants were able to recall correctly what they were told previous minorities / majorities said or did. Incongruencies between the given and recalled information were flagged. There was substantial heterogeneity with respect to the correct recall of these messages. We will return to this point in our additional data analysis in the second half of the results section.

norm-nudge content while keeping everything else constant shields us against experimenter demand effect as a credible explanation for observing differential behavioral responses.

Part 2: Lying Decision

Following the norm-nudge intervention, participants played 20 periods of the 'mind game' paradigm (Jiang, 2013).[7] Participants were first asked to think of a number between 1 and 6 and were then forwarded to a screen on which they rolled a digital die and observed the rolled number. A bonus ($0.10) would be paid only if participants claimed to have initially thought of the same number displayed on the digital die, with no bonus otherwise (in each round, we randomized the order in which these two options were displayed to the participants). Participants were paid the sum of all bonuses at the end of the experiment. Participants were made aware of the payoff rules at the beginning of the experiment. The experiment ended with a manipulation check (participants were shown all 5 possible norm-messages used in our experiment and were asked to correctly identify the message that they had been shown) and a summary of their payoffs.

2.3. Hypotheses

Prior research has shown that providing "norm-nudges", such as information on how relevant peers behave and what behavior they consider appropriate, can crucially affect whether individuals adhere to norms and, in turn, demonstrate behavioral change (Bicchieri and Dimant, 2019). Our paper seeks to understand how different versions of truthful norm-related information and their frames (positive versus negative) affect deviant behavior in the form of misreporting a random outcome for a monetary bonus. We differentiate between behaviors when "empirical information" of a norm (what other participants had previously done in a similar context with the option to lie) is provided and behavior when "normative information" is provided (what other participants said "is the appropriate thing to do" in a similar context). Frames are provided in either a positive (majority engaged in / approved of compliant behavior) or negative way (minority engaged in / approved of deviant behavior). We contrast behavioral results from these with behavior when neither empirical nor normative information is provided (Baseline).

[7] The experiment was played repeatedly to avoid a one shot high stakes decision. This verifies that participants would not shy away from lying because a lie would have (relatively) major consequences. Having multiple die rolls, participants may feel comfortable to lie to a limited extent, which is a common finding in this literature.

We model our hypotheses after the methodological discussion provided in Bicchieri and Dimant (2019). In their research, the authors provide a theoretical foundation of the existing social norms research, in particular with respect to nudging, and discuss its insights on the basis of other existing empirical research. We capitalize on these theoretical implications and investigate in our paper here the extent to which such norm-nudges can affect behavior and – ultimately – change social norms. With this in mind, we hypothesize that over-reporting, whereby subjects claim to have thought of the rolled die numbers at a rate higher than chance, will occur in our paradigm in the following order: Baseline > Normative > Empirical. We are able to derive this based on existing empirical literature discussed in Bicchieri and Dimant (2019), which shows that descriptive information can be more influential and informative than normative information in the context of trust and pro-social behavior (Bicchieri and Xiao, 2009; Bicchieri et al., 2019a,c). This can be attributed to the asymmetric signaling between descriptive and normative information in that providing the former ('walking the talk') allows for better inference of the latter than vice versa, which can merely be considered 'cheap talk' (e.g., Eriksson et al., 2015; Bicchieri et al., 2019b). Additionally, existing research (cited above) suggests that frames matter and individuals may react to positive and negative frames differently. However, the interaction of framing with norm information has yet to be explored and hence is an empirical question. We remain agnostic about the direction of the framing effect and predict the null.

2.4. Results

As specified in our pre-registration, our first analysis examines the average behavior across treatments using non-parametric tests in which we treat the average behavior of an individual across all 20 periods as one independent observation. Next, we utilize our repeated design and examine our results through the lens of a panel data analysis using a random-effects logit model with standard errors clustered at the individual level. Generally in line with existing literature, we observe reporting rates significantly greater than chance (16.67%), suggesting inflated reports through lying across all conditions. In both types of analyses, our data yielded highly powered and convincing null results with effect sizes hovering at around 0.1 (see Figures 2 and 3): using a Kruskal–Wallis test, we find no significant differences in average reporting behavior across all conditions (p=0.134).

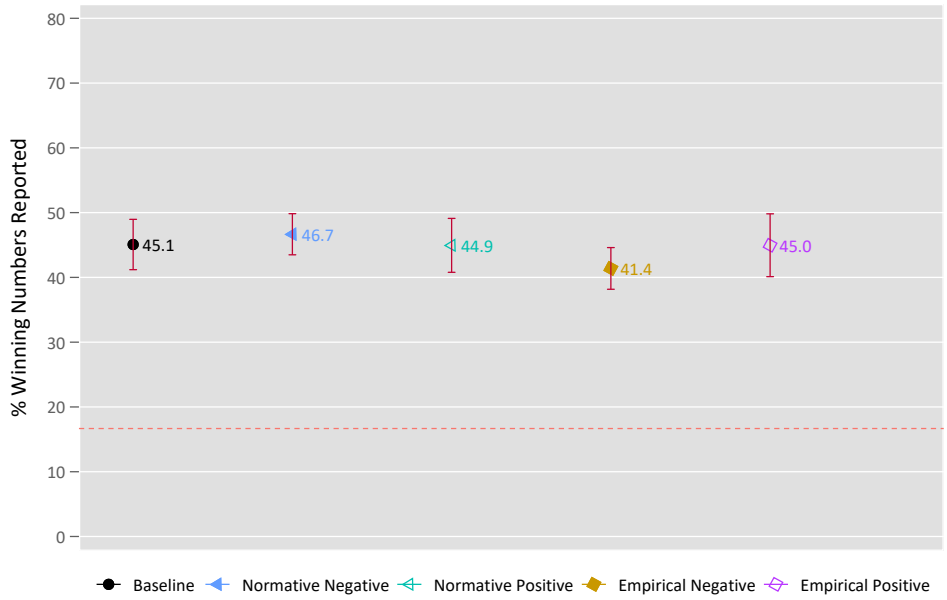

Figure 2: Mean reporting of winning numbers across all conditions. Red horizontal dashed line represents the expected value (16.67%) if all reports were truthful. Whiskers indicate 95% confidence intervals.

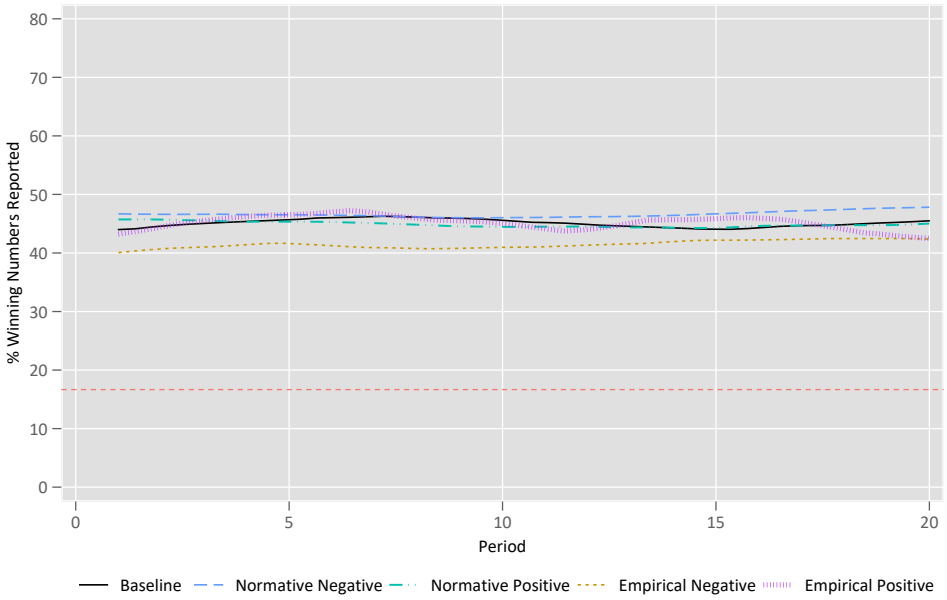

Figure 3: Mean reporting of winning numbers across periods and all conditions. Kernel-weighted local polynomial smoothing applied to lines for expository purposes. Red horizontal dashed line represents the expected value (16.67%) if all reports were truthful.

This is further corroborated by our results of pairwise mean comparisons that account for inflation of the type-I error through the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). We obtain consistently insignficant results for the pairwise comparisons, with the lowest p-value being 0.164 (Baseline vs. *Empirical Negative*). In fact, our confidence intervals are so small that no theoretically or economically meaningful effect size (based on the previously discussed literature) falls within its bounds.[8]

A similar picture arises for reporting decisions across periods. We do not observe much variation over time, nor any significant differences between treatments.[9] We examine this through the lens of a regression framework (Table 1). Consistent with the previous results, the treatment interventions yield no significant differences compared to the Baseline, even after controlling for relevant covariates including gender (reference group = male), period dummies, a participant's age, the SOEP risk measure (Wagner et al., 2007) (higher = more risk-seeking), and a control for whether participants met the criteria of post-hoc exclusion from the experiment, as specified in our pre-registration. We present exploratory analyses of these participants' behavior in the Appendix Figures A.2 and A.3.

In a last step, it is of particular importance to understand *why* such information-based nudges do not yield measurable behavioral change. For this, we follow the economic tradition of social norms research and run a follow-up study[10] (n=217, derived from the

[8]Exploratory, we provide more support for a true and reliable null effect by employing the TOST procedure as introduced by Lakens et al. (2018) to test for equivalence and reject the presence of a smallest effect size of interest, with the p-values being 0.031 or smaller, thus further strengthening the null-effects.

[9]It is important to rule out that the observed null results are due to misunderstandings of the norm-nudge information, e.g. because a positively framed message ('... a vast majority lies / approves of lying') might be more intuitive than a negatively framed message ('...a vast minority does not lies / does not approve of lying'). We address this concern by running a follow-up experiment on MTurk with the same show-up fee ($0.30) and data quality restrictions as used in our main experiment. We collected beliefs from n=250 participants who previously have not participated in our original experiment and presented them with one random original norm-message (same as previously observed by the participants in our main experiment). Depending on the treatment, participants were incentivized to guess the fraction of participants who either lied or approved of lying in the initial experiment from which this information was truthfully derived. Participants received a monetary bonus ($0.30) if they provided a guess that fell within 10% of the correct fraction. Screenshots of the experiment are available upon request. Those who read the positive (negative) empirical information guessed that the fraction of participants that lied (did not lie) was 81.3% (39.1%), which is significantly above (below) 50%. Similarly, those who read the positive (negative) normative information, participants guessed that the fraction of participants that said one should be honest (should not be honest) was 78.6% (41.7%), which again is significantly above (below) 50%. See Appendix Figure A.1 for an illustration. These results allow us to conclude that our original null results cannot be explained by potential confusion on the side of the participants to properly process the norm information.

[10]As before, we pay a show-up fee of $0.30 and apply the same quality restrictions on the data as in our

DV: % Reporting Winning Number	(1)	(2)	(3)
Treatment (Base level: Baseline)			
Normative Negative (NN)	1.192 (0.172)	0.871 (0.439)	1.014 (0.159)
Normative Positive (NP)	1.005 (0.162)	0.532 (0.290)	0.965 (0.157)
Empirical Negative (EN)	0.873 (0.125)	0.680 (0.343)	0.763* (0.117)
Empirical Positive (EP)	1.071 (0.194)	1.064 (0.671)	1.084 (0.196)
Female	0.859 (0.085)	0.708 (0.157)	0.857 (0.085)
Female × NN		1.210 (0.353)	
Female × NP		1.531 (0.498)	
Female × EN		1.171 (0.344)	
Female × EP		0.978 (0.358)	
Period	1.003 (0.002)	1.003 (0.002)	1.003 (0.002)
Age	0.966*** (0.004)	0.967*** (0.004)	0.967*** (0.004)
Risk	1.086*** (0.023)	1.086*** (0.023)	1.081*** (0.022)
Exclusion			1.327** (0.149)
Observations	24000	23840	24000

Table 1: Random-effects logit regressions. Coefficients denote odds ratios. Standard errors in parentheses and are clustered at the individual level. Stars indicate significant differences at the conventional levels of *p<0.1, **p<0.05, and ***p<0.01.

power-calculation of our main experiment to achieve a power of at least 80%) in which we test the assumption that the mere presence of norm-information might have been unable to actually *shift* an existing norm using the norm elicitation method as introduced by Krupka and Weber (2013). This line of thought is consistent with the assumption that provision of information is often not enough as behaviors can be sticky. Instead, one needs more than just a gentle nudge – for example a 'shove' – to achieve actual behavioral change (e.g., Kahan, 2000; Oliver, 2015; Grüne-Yanoff and Hertwig, 2016; Loewenstein and

previous experiments. Screenshots of the experiment are available upon request.

Chater, 2017; Ariely, 2019). This is indeed what we find (see Figure 4): across the different norm-information conditions, participants are asked to rate the (in)appropriateness of lying behavior after internalizing the respective message. Convincingly, none of the norm elicitation results yield any significant changes in norm perceptions compared to the Baseline: p-values (left-to-right) from $\tilde{\chi}^2$ tests correspond to 0.19, 0.25, 0.17, and 0.49, respectively.

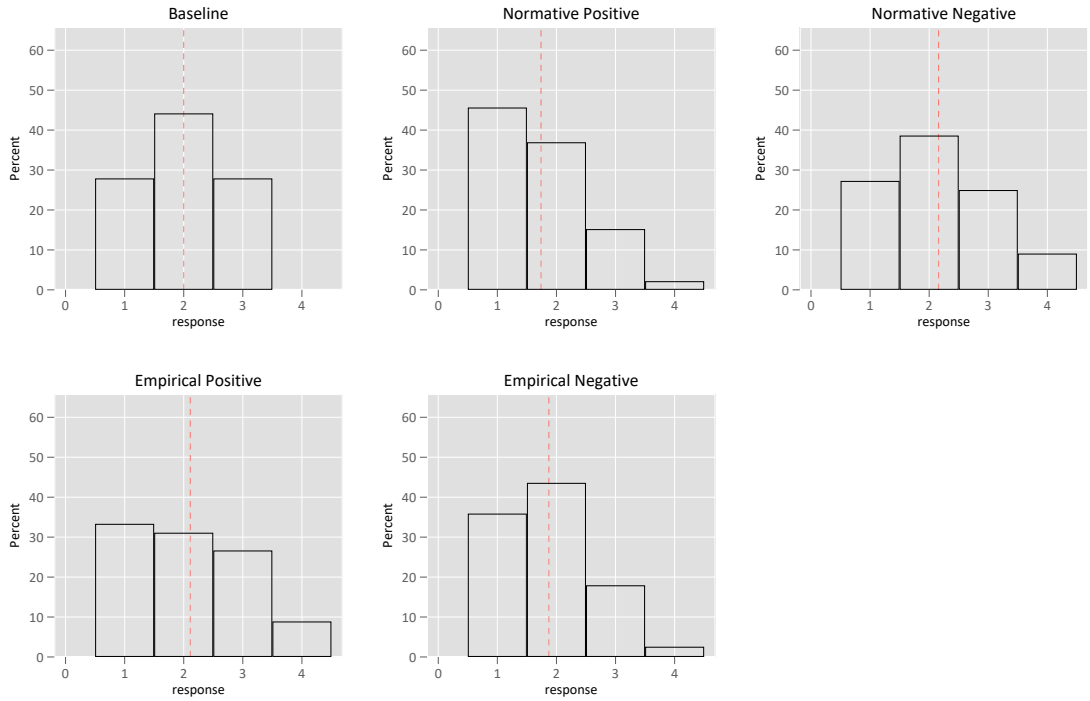

Figure 4: Measuring norms using the elicitation method as introduced by Krupka and Weber (2013). Values of 1, 2, 3, and 4 correspond to the normative assessments 'Very Socially Inappropriate', 'Somewhat Socially Inappropriate', 'Somewhat Socially Appropriate', and 'Very Socially Appropriate', respectively. Vertical red dotted lines represent averages.

3. Conclusion and Discussion

Reducing deviant behavior and increasing normative and ethically desirable behavior is an important goal. We attempted to achieve this by subtly nudging participants to consider what the social norm is among their fellow participants. Specifically, we informed participants about what other participants *previously did* (empirical message) or *approved of doing* (normative message). We further varied whether the norm was framed in a positive way (majority did not cheat / did not approve of cheating) or an inverted and

equivalent negative way (minority cheated / approved of cheating). Participants then had an opportunity to increase their payoffs by lying about predicted die roll numbers. Results revealed robust support for the null hypothesis, suggesting the way the information was presented did not affect participants' level of (dis)honesty.

Lack of support for differences between experimental treatments does not necessarily mean supporting the null. Several factors, however, speak to the fact that the null is, in our case, more likely than any alternative possibility. First, we have calculated, pre-registered, and collected data based on a power-analysis that allowed us to detect even small effect sizes at the conventional significance thresholds. Including our additional, pre-registered data collection, we achieve statistical power in excess of 95%. Our results, however, indicate convincingly against statistically or economically relevant effect sizes across all treatments and specifications. Second, additional checks employing a Bayesian analysis and the TOST procedure (Lakens et al., 2018) revealed that the null hypothesis is substantially more likely than the alternative hypothesis suggesting a difference in dishonesty between treatments. Third, we obtain suggestive evidence for a 'convenient misperception' of the content of these norm-nudges among a sizeable subset of participants. For these participants, we observe a higher rate of lying overall, which is in line with the literature on motivated misremembering and belief distortion, among others (e.g., Carlson et al., 2018; Exley and Kessler, 2018; Gneezy et al., 2018b; Zimmermann, 2019; Bicchieri et al., 2019b; Saucet and Villeval, 2019). Lastly, we find that one reason for the overall ineffectiveness of our norm-nudge interventions is the inability to actually *shift* the perception of existing norms as measured by Krupka and Weber (2013).

The results obtained in the current study are valuable in informing social norm theorizing. Specifically, recent work – particularly in the field – has revealed that reminding people about descriptive and injunctive norms can lead to a positive behavioral response, especially when implemented with care and regard for the cultural and institutional circumstances (e.g., Hallsworth et al., 2017; Hernandez et al., 2017; Bott et al., 2019, and see also discussion in Bicchieri and Dimant, 2019). Importantly, the literature has also produced conflicting results in that not all interventions are equally successful, as some interventions have been observed to yield no effect or even a backfiring effect (e.g., Blumenthal et al., 2001; Fellner et al., 2013; Bicchieri et al., 2019c).

Clearly, obtaining a better understanding of the settings in which nudging good behavior works is practically important. Policy makers seek to use the most effective interventions and tailor those to the settings in which they operate, though our results cast doubt on

the effectiveness of negatively framed norm-messages. To increase good behavior, stronger interventions should be considered that involve social and/or economic incentives (see, e.g., Bolton et al., 2019), which should be studied both in the lab and in the field to address challenges with respect to disentangling the underlying mechanisms for deviant behavior using this and related cheating paradigms (see extensive discussions and results presented in, e.g., Hao and Houser, 2017; Abeler et al., 2019; Gerlach et al., 2019; Köbis et al., 2019).

Future research could extend our insights along two domains. While we find that simple norm-nudges can be unsuccessful at shifting norms in environments in which a behavioral norm is already in place (cheating is normatively not acceptable, as our results from the Krupka and Weber (2013) elicitation suggest), it remains an open question whether such gentle interventions are more effective when norms are not clearly defined or known. Arguably, soft interventions in such environments are tricky because they give rise to self-serving belief manipulation (for theoretical and experimental evidence see Bicchieri et al., 2019b), which in turn may undermine the effectiveness of norm-nudges. The extent to which the effectiveness of norm-interventions is mediated by the common perception of an existing norm is subject to scientific debate (see, e.g., Bicchieri et al., 2019c) and requires further research. In addition, while our results can speak to the effectiveness of such interventions when the risk of detection for deviant behavior is absent by design (in our case implemented through the mind game as per Jiang, 2013), the results may or may not be comparable when risk of detection is present (as is often the case in the standard dice paradigm of Fischbacher and Föllmi-Heusi, 2013). This would be in line with recent literature that has established important differences between paradigms in which lying can and cannot be observed, such as responsiveness in cheating behavior to incentives (Kajackaite and Gneezy, 2017). For this reason, norm-nudge interventions as employed in our context could be more effective when risk of detection, or at least the observability of one's actions (as, for example, in the context of the well-known hotel towel study by Goldstein et al., 2008) is present. The working assumption is that such information-based interventions can trigger changes in perceptions of the likelihood of being caught, suggesting that, for example, a government that is sending personalized letters may be 'onto something' (Chalfin and McCrary, 2017; Bott et al., 2019).

References

Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for truth-telling. *Econometrica*.

Alekseev, A., Charness, G., and Gneezy, U. (2017). Experimental methods: When and why contextual instructions are important. *Journal of Economic Behavior & Organization*, 134:48–59.

Ambuehl, S., Bernheim, B. D., and Ockenfels, A. (2019). Projective paternalism. Working Paper 26119, National Bureau of Economic Research.

Andreoni, J. (1995). Warm-glow versus cold-prickle: the effects of positive and negative framing on cooperation in experiments. *The Quarterly Journal of Economics*, 110(1):1–21.

Andreoni, J. and Bernheim, B. D. (2009). Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77(5):1607–1636.

Arechar, A. A., Gächter, S., and Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1):99–131.

Ariely, D. (2019). How to change your behavior for the better. Ted salon talk, available at: https://tinyurl.com/arielyted2019.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American economic review*, 96(5):1652–1678.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C. and Dimant, E. (2019). Nudging with Care: The Risks and Benefits of Social Information. *Public Choice*.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2019a). Observability, social proximity, and the erosion of norm compliance. Working Paper Available at SSRN: https://ssrn.com/abstract=3355028.

Bicchieri, C., Dimant, E., and Sonderegger, S. (2019b). It's not a lie if you believe the norm does not apply: Conditional norm-following with strategic beliefs. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3326146.

Bicchieri, C., Dimant, E., and Xiao, E. (2019c). Deviant or wrong? the effects of norm information on the efficacy of punishment. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294371.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Blumenthal, M., Christian, C., Slemrod, J., and Smith, M. G. (2001). Do normative appeals affect tax compliance? evidence from a controlled experiment in minnesota. *National Tax Journal*, pages 125–138.

Bolton, G., Dimant, E., and Schmidt, U. (2019). The fragility of nudging: Combining (im)plausible deniability with social and economic incentives to promote behavioral change. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3294375.

Bott, K. M., Cappelen, A. W., Sorensen, E., and Tungodden, B. (2019). You've got mail: A randomised field experiment on tax evasion. *Management Science*.

Boyer, P. C., Dwenger, N., and Rincke, J. (2016). Do norms on contribution behavior affect intrinsic motivation? field-experimental evidence from germany. *Journal of Public Economics*, 144:140–153.

Brandon, A., Ferraro, P. J., List, J. A., Metcalfe, R. D., Price, M. K., and Rundhammer, F. (2017). Do the effects of social nudges persist? theory and evidence from 38 natural field experiments. Technical report, National Bureau of Economic Research.

Buckenmaier, J., Dimant, E., Posten, A.-C., and Schmidt, U. (2019). Efficient institutions and effective deterrence: On timing and uncertainty of punishment. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.3300563.

Bursztyn, L., Fiorin, S., Gottlieb, D., and Kanz, M. (2019). Moral incentives in credit card debt repayment: Evidence from a field experiment. *Journal of Political Economy*, 127(4):000–000.

Carlson, R. W., Marechal, M., Oud, B., Fehr, E., and Crockett, M. (2018). Motivated misremembering: Selfish decisions are more generous in hindsight. Working paper.

Castro, L. and Scartascini, C. (2015). Tax compliance and enforcement in the pampas evidence from a field experiment. *Journal of Economic Behavior & Organization*, 116:65–82.

Chalfin, A. and McCrary, J. (2017). Criminal Deterrence: A Review of the Literature. *Journal of Economic Literature*, 55(1):5–48.

Chang, D., Chen, R., and Krupka, E. (2019). Rhetoric matters: A social norms explanation for the anomaly of framing. *Games and Economic Behavior*, 116:158–178.

Charness, G., Blanco-Jimenez, C., Ezquerra, L., and Rodriguez-Lara, I. (2019). Cheating, incentives, and money manipulation. *Experimental Economics*, 22(1):155–177.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., and Winter, P. L. (2006). Managing social norms for persuasive impact. *Social influence*, 1(1):3–15.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621.

Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Coppock, A., Leeper, T. J., and Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49):12441–12446.

Cranor, T., Goldin, J., Homonoff, T., and Moore, L. (2018). Communicating tax penalties to delinquent taxpayers: Evidence from a field experiment. Working paper.

Deutsch, M. and Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3):629.

Dimant, E. (2019). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., Nellis, G., Adida, C. L., Arias, E., Bicalho, C., Boas, T. C., et al. (2019). Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances*, 5(7):eaaw2612.

Dupas, P. (2011). Do teenagers respond to hiv risk information? evidence from a field experiment in kenya. *American Economic Journal: Applied Economics*, 3(1):1–34.

Eriksson, K., Strimling, P., and Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, 129:59–69.

Exley, C. and Kessler, J. B. (2018). Motivated errors. *Working Paper*.

Fellner, G., Sausgruber, R., and Traxler, C. (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 Föllmi-Heusi, F. (2013). Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.

Gächter, S., Orzen, H., Renner, E., and Starmer, C. (2009). Are experimental economists prone to framing effects? a natural field experiment. *Journal of Economic Behavior & Organization*, 70(3):443–446.

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

Gino, F., Ayal, S., and Ariely, D. (2009). Contagion and differentiation in unethical behavior: The effect of one bad apple on the barrel. *Psychological science*, 20(3):393–398.

Gino, F., Hauser, O. P., and Norton, M. I. (2019). Budging beliefs, nudging behaviour. *Mind & Society*, pages 1–12.

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

Gneezy, U., Saccardo, S., and van Veldhuizen, R. (2018b). Bribery: Behavioral drivers of distorted decisions. *Journal of the European Economic Association*, 17(3):917–946.

Goldstein, N. J., Cialdini, R. B., and Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, 35(3):472–482.

Grüne-Yanoff, T. and Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, 26(1-2):149–183.

Hallsworth, M., List, J. A., Metcalfe, R. D., and Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148:14–31.

Hao, L. and Houser, D. (2017). Perceptions, intentions, and cheating. *Journal of Economic Behavior & Organization*, 133:52–73.

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., and Bigham, J. P. (2018). A data-driven analysis of workers' earnings on amazon mechanical turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 449. ACM.

Hernandez, M., Jamison, J., Korczyc, E., Mazar, N., and Sormani, R. (2017). *Applying Behavioral Insights to Improve Tax Collection*. World Bank.

Hildreth, J. A. D., Gino, F., and Bazerman, M. (2016). Blind loyalty? when group loyalty makes us see evil or engage in it. *Organizational Behavior and Human Decision Processes*, 132:16–36.

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

John, P. and Blume, T. (2018). How best to nudge taxpayers? the impact of message simplification and descriptive social norms on payment rates in a central london local authority. *Journal of Behavioral Public Administration*, 1(1).

Kahan, D. M. (2000). Gentle nudges vs. hard shoves: Solving the sticky norms problem. *The University of Chicago Law Review*, pages 607–645.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, 58(9):697.

Kajackaite, A. (2018). Lying about luck versus lying about performance. *Journal of Economic Behavior & Organization*, 153:194–199.

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

Kettle, S., Hernandez, M., Sanders, M., Hauser, O., and Ruda, S. (2017). Failure to captcha attention: Null results from an honesty priming experiment in guatemala. *Behavioral Sciences*, 7(2):28.

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science*, 14(5):778–796.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Lakens, D., Scheel, A. M., and Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

Larkin, C., Sanders, M., Andresen, I., and Algate, F. (2019). Testing local descriptive norms and salience of enforcement action: A field experiment to increase tax collection. *Journal of Behavioral Public Administration*, 2(1):1–11.

Lee, M. D. and Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge university press.

Levin, I. P., Schneider, S. L., and Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, 76(2):149–188.

Loewenstein, G. and Chater, N. (2017). Putting nudges in perspective. *Behavioural Public Policy*, 1(1):26–53.

Luttmer, E. F. and Singhal, M. (2014). Tax morale. *Journal of Economic Perspectives*, 28(4):149–68.

Mazar, N., Amir, O., and Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research*, 45(6):633–644.

Oliver, A. (2015). Nudging, shoving, and budging: Behavioural economic-informed policy. *Public Administration*, 93(3):700–714.

Peer, E., Acquisti, A., and Shalvi, S. (2014). "i cheated, but only a little": Partial confessions to unethical behavior. *Journal of Personality and Social Psychology*, 106(2):202.

Reinikka, R. and Svensson, J. (2005). Fighting corruption to improve schooling: Evidence from a newspaper campaign in uganda. *Journal of the European economic association*, 3(2-3):259–267.

Rogers, T. and Feller, A. (2018). Reducing student absences at scale by targeting parents' misbeliefs. *Nature Human Behaviour*, 2(5):335.

Saucet, C. and Villeval, M. C. (2019). Motivated memory in dictator games. *Games and Economic Behavior*.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological science*, 18(5):429–434.

Shalvi, S., Dana, J., Handgraaf, M. J., and De Dreu, C. K. (2011). Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior. *Organizational Behavior and Human Decision Processes*, 115(2):181–190.

Snowberg, E. and Yariv, L. (2018). Testing the waters: Behavior across participant pools. Working Paper 24781, National Bureau of Economic Research.

Sonnemans, J., Schram, A., and Offerman, T. (1998). Public good provision and public bad prevention: The effect of framing. *Journal of Economic Behavior & Organization*, 34(1):143–161.

Sunstein, C. R. (2017). Nudges that fail. *Behavioural Public Policy*, 1(1):4–25.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

van Kleef, G. A., Gelfand, M. J., and Jetten, J. (2019). The dynamic nature of social norms: New perspectives on norm development, impact, violation, and enforcement.

Wagner, G. G., Frick, J. R., and Schupp, J. (2007). The german socio-economic panel study (soep)-evolution, scope and enhancements. Report.

Zimmermann, F. (2019). The dynamics of motivated beliefs. Forthcoming American Economic Review.

Appendix A. Appendix

Appendix A.1. Robustness Checks

We ran a follow-up experiment in order to ensure that our main findings cannot be explained by potential confusion on the side of the participants in how the provided norm information was interpreted. As detailed in Footnote 9, we ran a follow-up study (n=250 in total who did not participate in the experiment, see legend of Figure A.1 for detailed breakdown per treatment) that was informed by the same power calculations as in our main experiment (effect size of 0.387 to achieve 80% for a signrank test). Restrictions to the MTurk sample selection were the same as before (see Footnote 3 for more details). The procedure was straightforward: we presented participants one of four norm message combinations (positive/negative framing × empirical/normative message) at random and asked them to guess in an incentive-compatible way the fraction of participants engaging in those behaviors (as per our collected data). Our results reliably indicate that positive (negative) messages are perceived as significantly larger (smaller) than 50%.

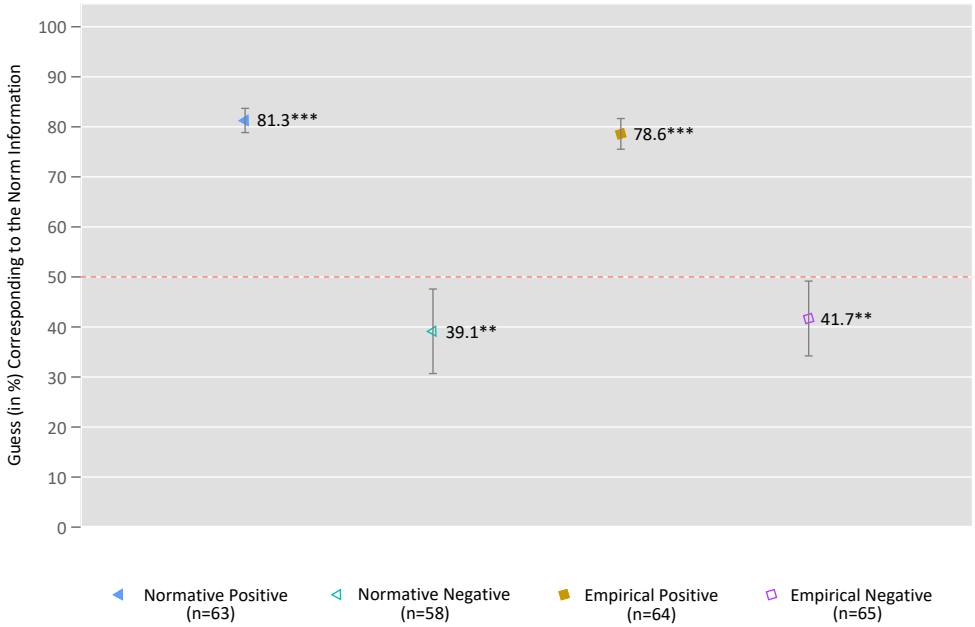

Figure A.1: Robustness check to examine the comprehension of the original norm-nudge messages. Positive information, which refers to majority behavior, is expected to be above 50%, whereas negative information, which refers to minority behavior, is expected to be below 50%. Stars next to point estimates indicate significant differences at the conventional levels (*p<0.1, **p<0.05, and ***p<0.01) from the 50% threshold. Whiskers indicate 95% confidence intervals.

Beyond our conclusive main null results as presented in Section 2.4, one particular result deserves more attention, which we will treat in the form of a secondary analysis. Recall that we included several attention and manipulation checks to retain control over the validity of our treatment interventions. Following the presentation of the norm-nudge information, participants were asked to recall and write up this information in as much detail as possible on the next screen. Failure to recall the information correctly was coded, which in turn allows us to examine the behavioral differences between those who passed and those who failed the attention check.[11] Recent experimental literature on selective memory, biased recall, motivated beliefs, and belief distortion (e.g., Carlson et al., 2018; Exley and Kessler, 2018; Gneezy et al., 2018b; Zimmermann, 2019; Bicchieri et al., 2019b; Saucet and Villeval, 2019) suggests that a participant's relationship with actual behavior is malleable– often in self-serving ways. With this in mind, we pre-registered the plausible assumption that there is similar scope for such distortions in the context of norm-nudge provisions, which may carry over to behavioral differences within our paradigm.

Before turning to the between-group comparison, we first substantiate our original null result further by performing a Bayesian analysis of the contingency table of reported outcomes by treatment and exclusion. In line with our previous analysis of the complete data set, the size of the Bayes Factor (BF_{01}=508.319) reveals *decisive evidence* (Lee and Wagenmakers, 2014) for the null hypothesis of no treatment difference for the subset of participants who passed the attention check. Next, as can be inferred from the third model specification of Table 1, our results are in line with the assumption that a wrong recall can result in different behavior. Indeed, those who did not pass the attention check (and thus were put in the 'Exclusion' category) were about 33% more likely to report a winning number than those who did pass the attention check (p=0.012). This is also supported by a non-parametric Kruskal-Wallis test, suggesting that the behavior of those participants who were flagged is significantly different from those who correctly recalled the norm-nudge (p=0.0348). Paralleling our previous analysis, we obtain a Bayes Factor (BF_{01}) of nearly 0 when comparing these two groups, yielding decisive evidence for the alternative hypotheses of significant behavioral difference between these two groups. Figures A.2 and A.3 detail the means and changes across periods. Figure A.4 reports distributions across treatments,

[11] A research assistant who was blind to the design and research goals of this project was asked to code the statements. For messages to be counted as correct, participants had to at least correctly recall the basic scaffold of the presented norm-nudge.

none of which are significant against the Baseline at the 5% level ($\tilde{\chi}^2$ tests).

A more detailed text-based analysis of the responses reveals an interesting pattern: participants over-proportionally misremembered the messages in a 'convenient' way that completely changed the content of the norm-nudge (e.g., from "The vast **minority** [...] were dishonest..." to "The vast **majority** [...] were dishonest..."). We take this differential effect as further evidence that participants perceived and understood the messages as intended, since otherwise one would expect a uniform effect across all treatments. Although these results mirror existing literature, we remain agnostic about the exact mechanism yielding a 'convenient' and incorrect recall of norm-nudges in our paradigm.[12]

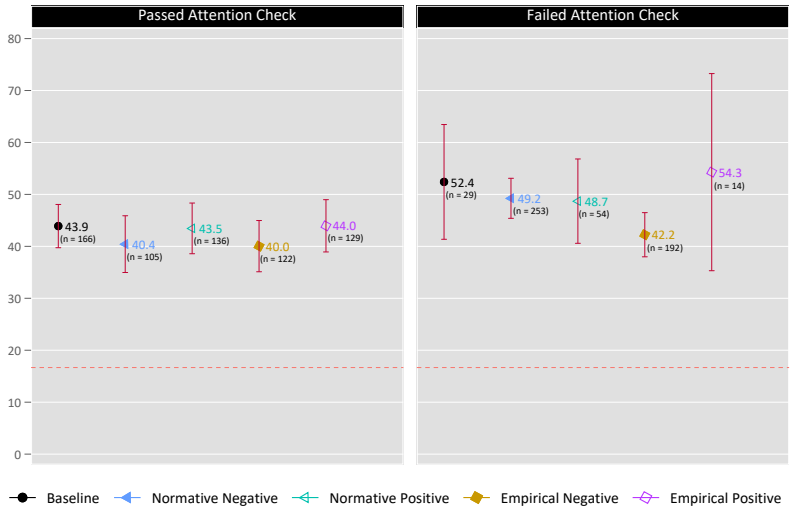

Figure A.2: Differentiation between participants who passed and failed the attention check. Mean reporting of winning numbers across all conditions. Red horizontal dashed line represents the expected value (16.67%) if all reports were truthful.

[12] This surprising result was the reason for our second pre-registration and the need to collect much more data than initially anticipated. For more details see pre-registration documents.

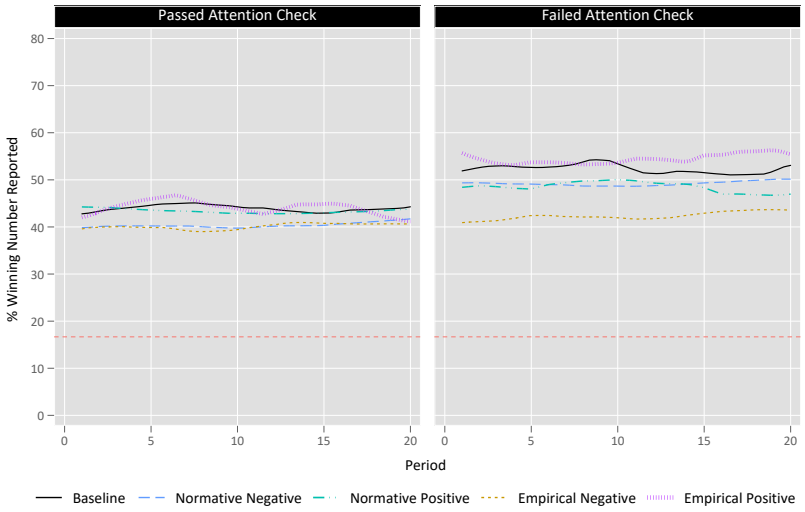

Figure A.3: Differentiation between participants who passed and failed the attention check. Mean reporting of winning numbers across periods and all conditions. Kernel-weighted local polynomial smoothing applied to lines for expository purposes. Red horizontal dashed line represents the expected value (16.67%) if all reports were truthful.

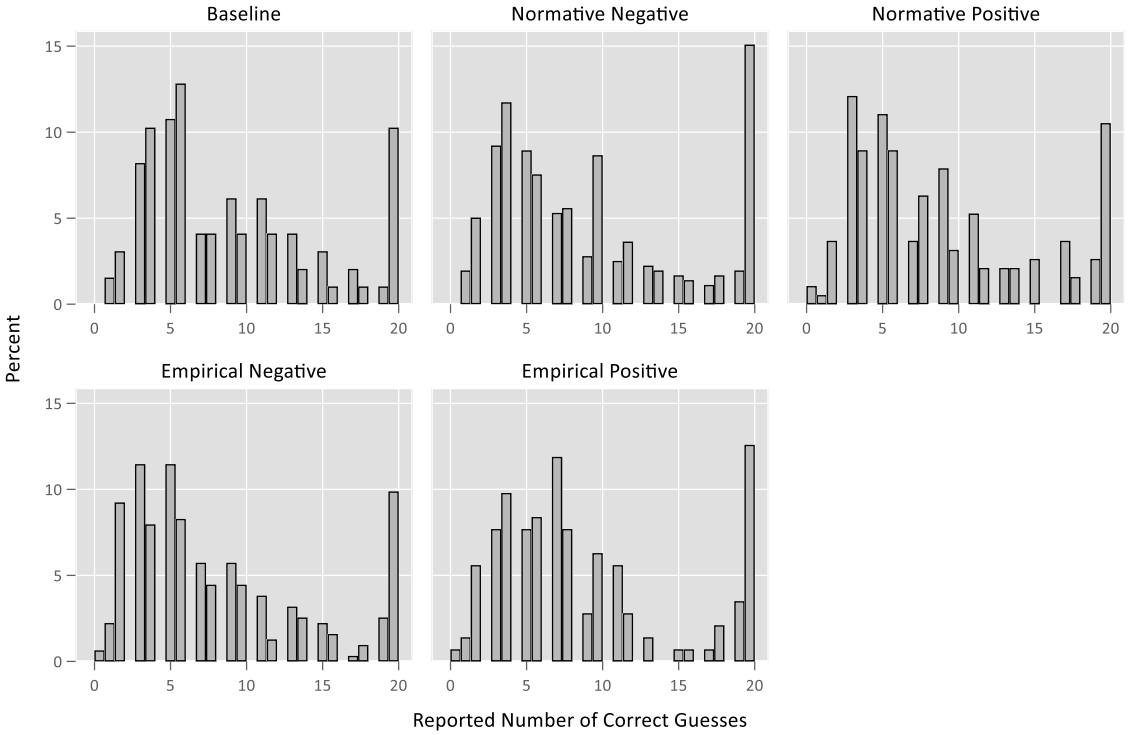

Figure A.4: Histogram of reports across treatments

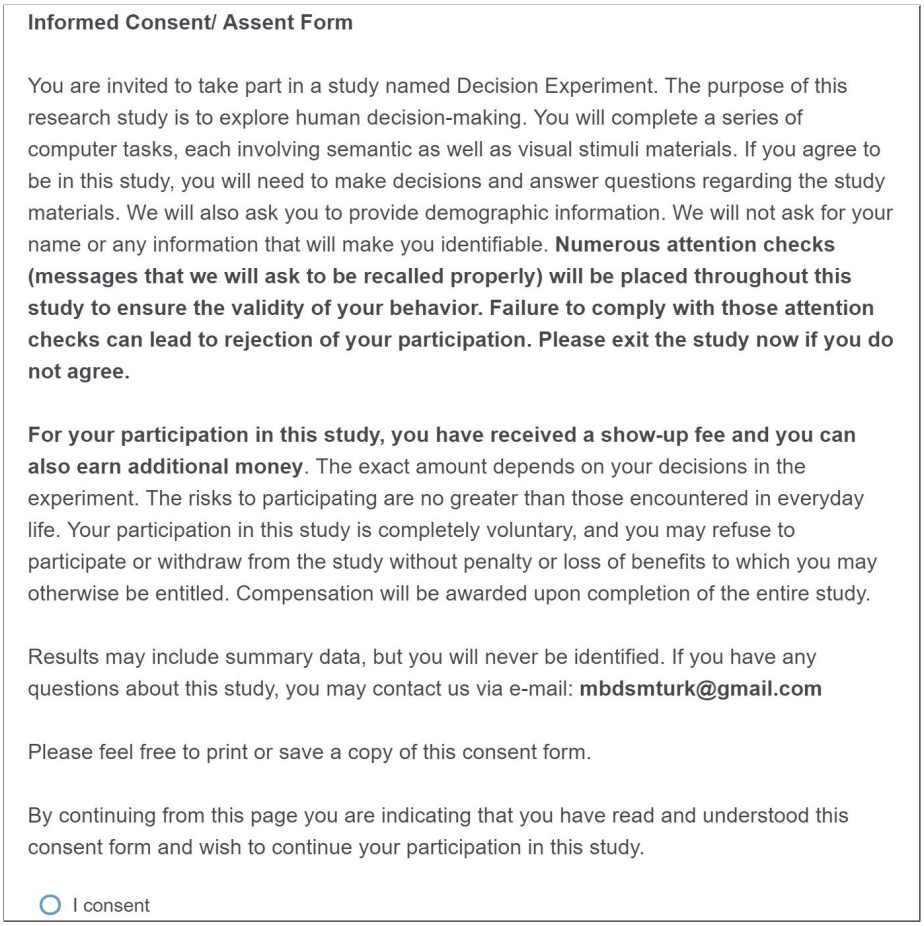

Figure A.5: Consent form

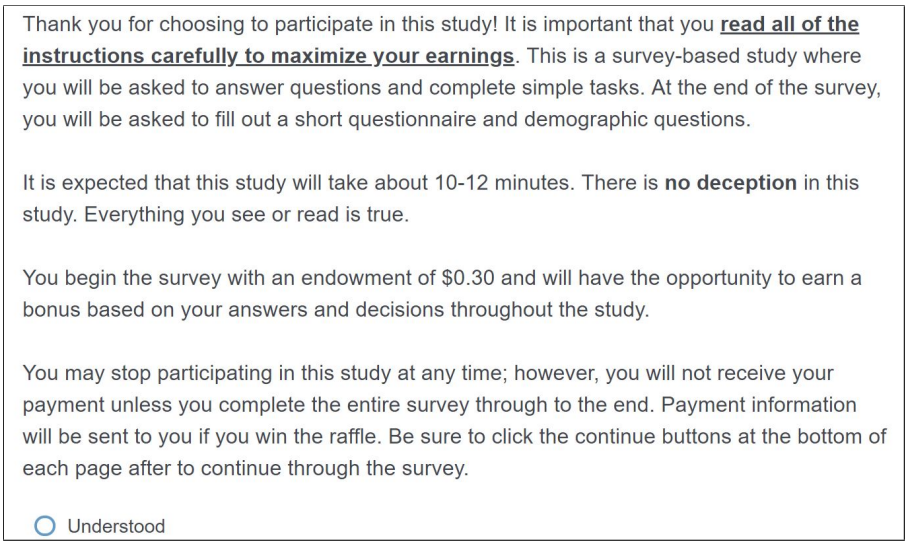

Figure A.6: Instructions

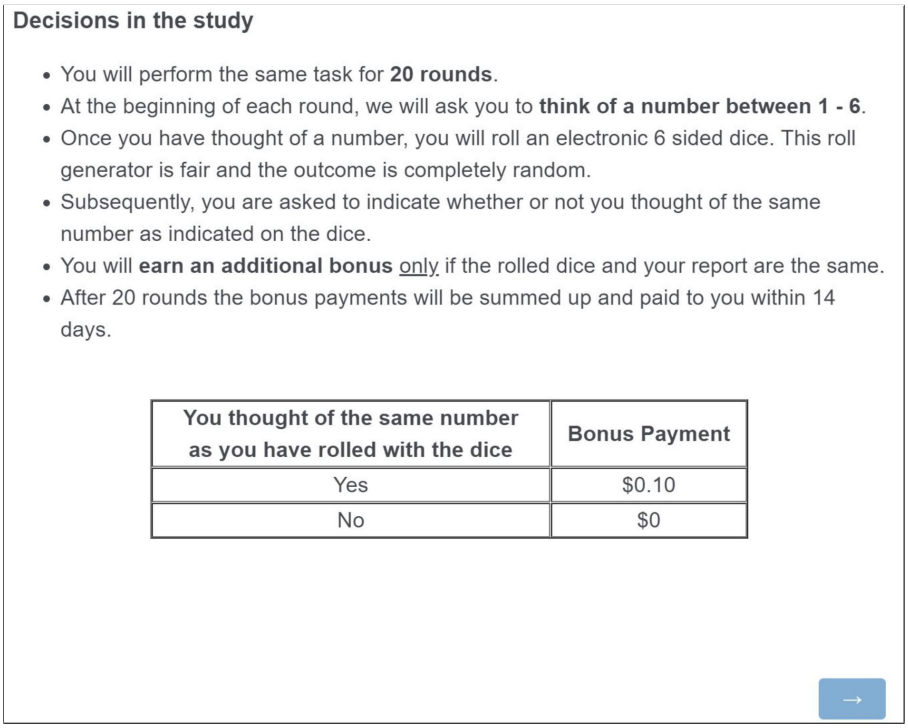

Figure A.7: Instructions

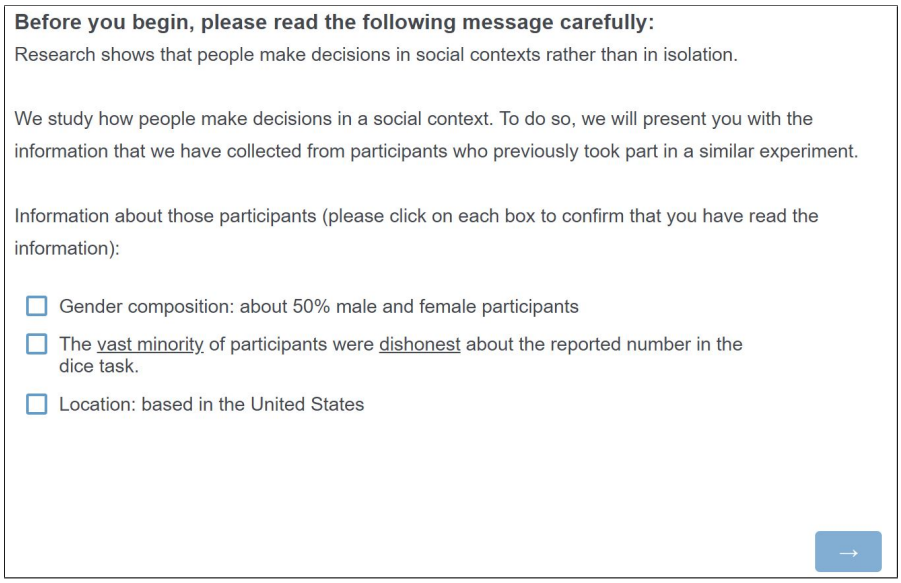

Figure A.8: Norm-Nudge (exemplary for the empirical negative condition

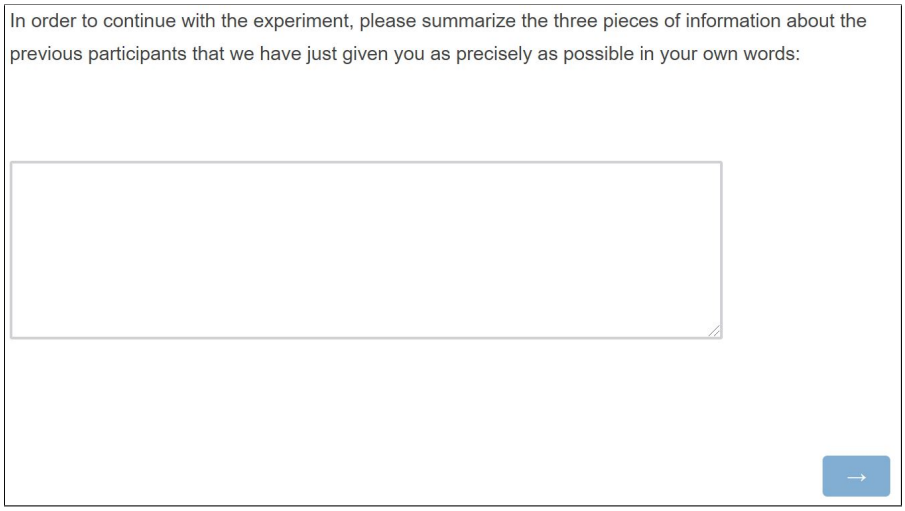

Figure A.9: Attention check

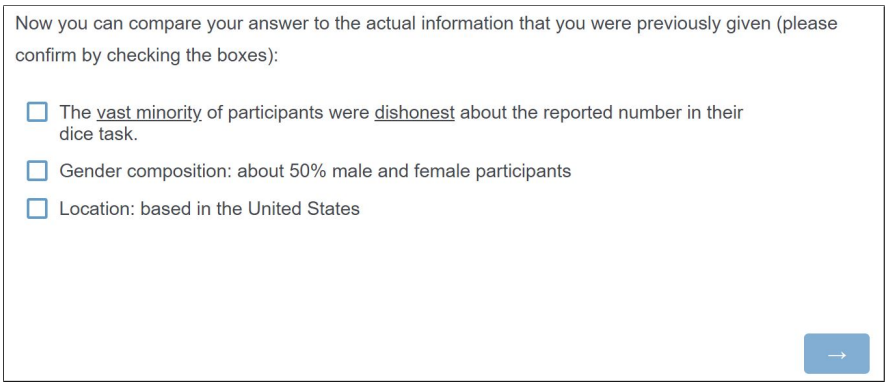

Figure A.10: Attention check with correct solution

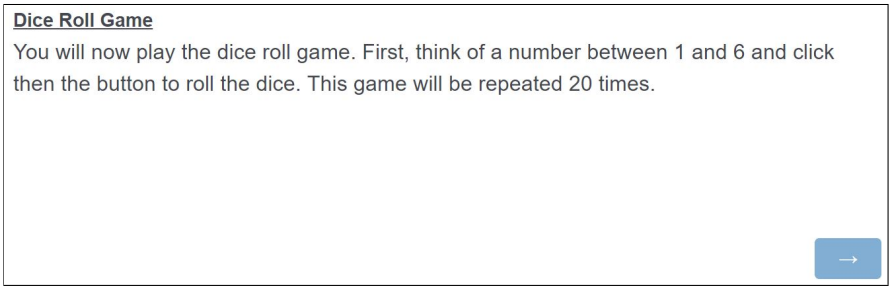

Figure A.11: Dice roll (first out of 20)

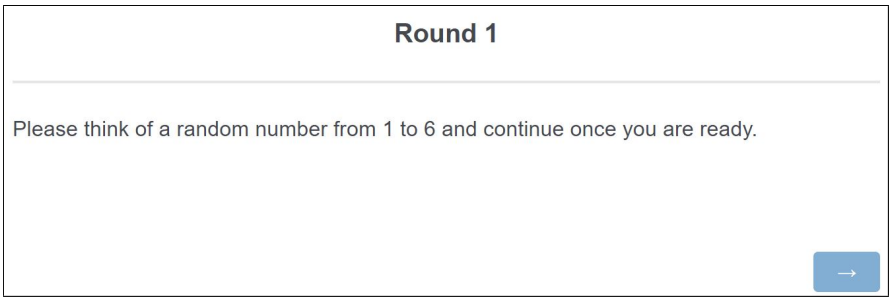

Figure A.12: Dice roll (first out of 20)

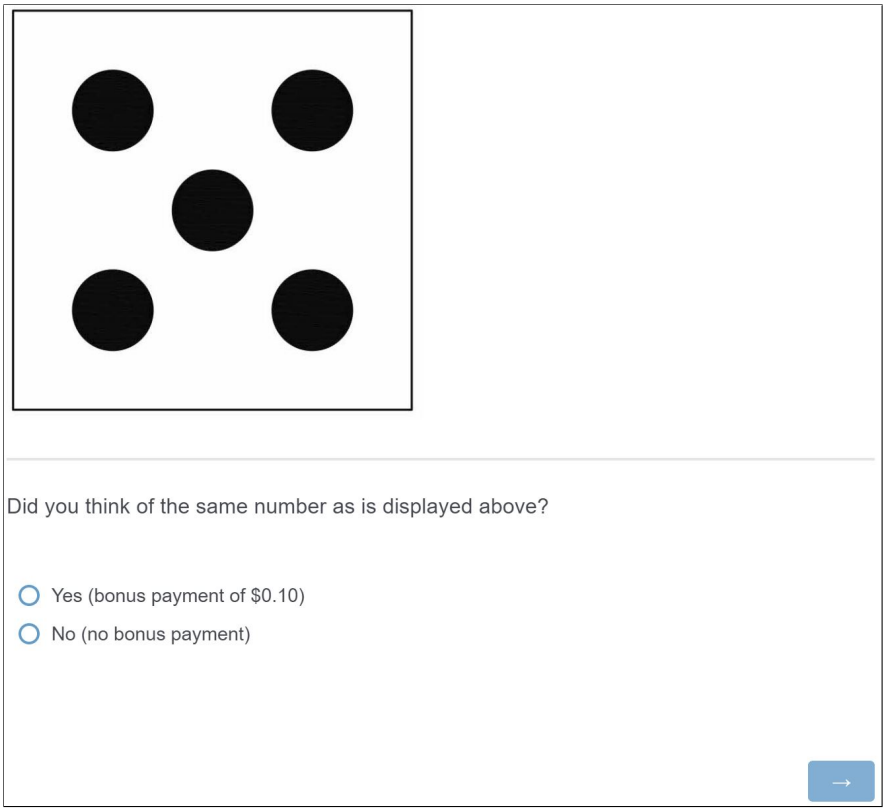

Figure A.13: Dice roll (first out of 20)

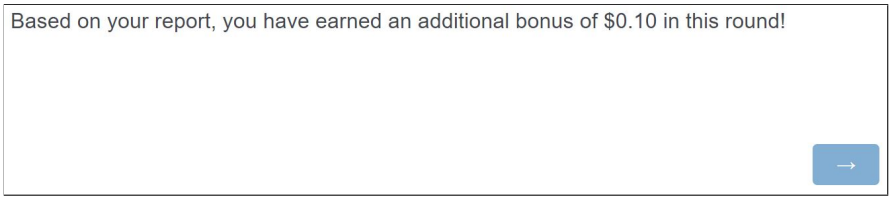

Figure A.14: Dice roll (first out of 20)

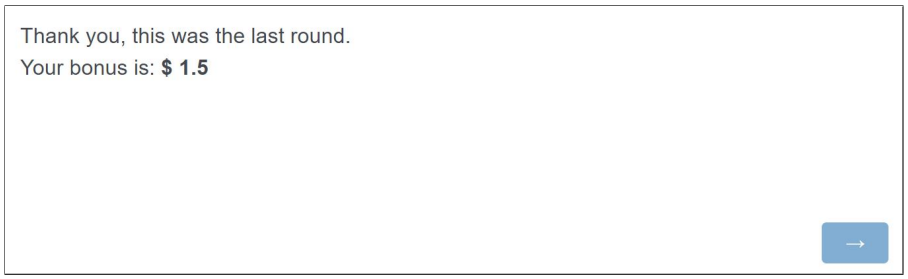

Figure A.15: Total payment screen at the end of 20 rounds

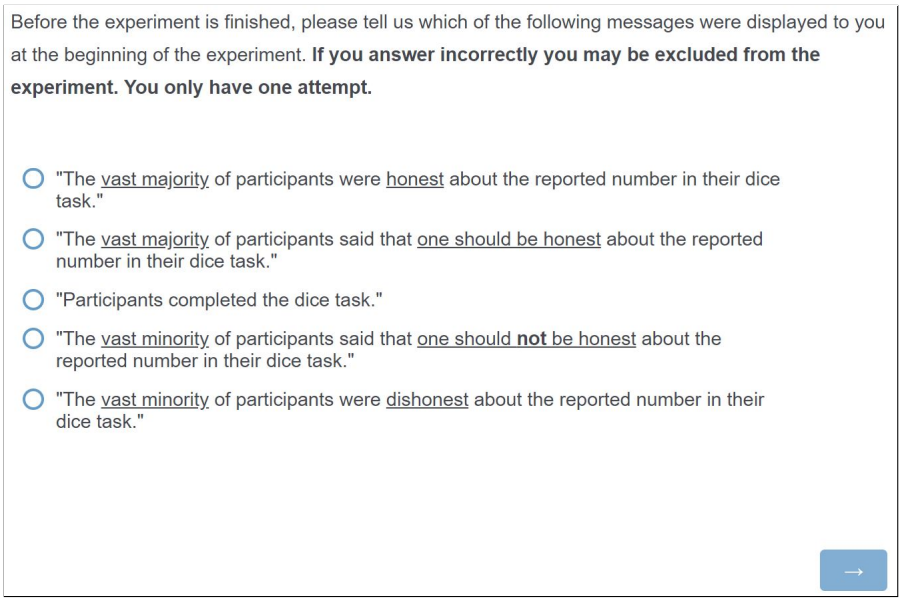

Figure A.16: Manipulation check of initial norm-nudge

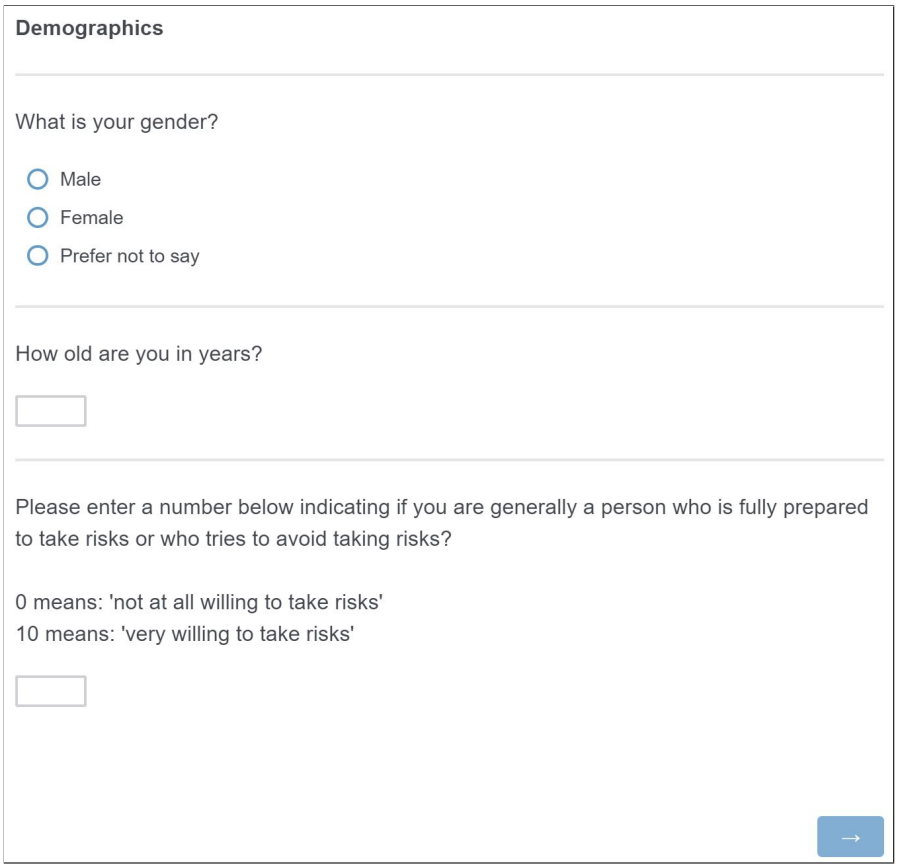

Figure A.17: Demographics

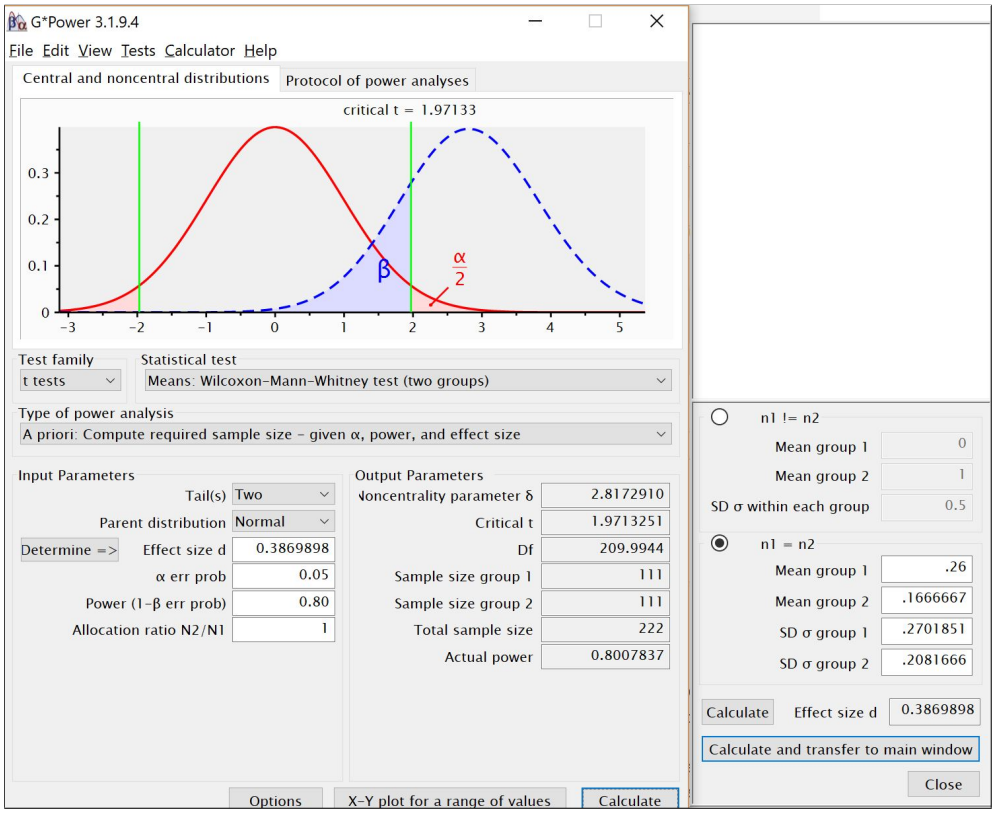

Figure A.18: Power Calculations for Pre-Registration