Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Case Report

Are implicit attitudes toward dishonesty associated with self-serving dishonesty? Implications for the reliability of the IAT

Hirokazu Hatta [a], Ryuhei Ueda [b,c], Hiroshi Ashida [a], Nobuhito Abe [b,*]

[a] *Graduate School of Letters, Kyoto University, Kyoto, Japan*
[b] *Kokoro Research Center, Kyoto University, Kyoto, Japan*
[c] *Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Osaka, Japan*

ARTICLE INFO

Editor: Rachel Barkan

Keywords:
Dishonesty
Honesty
Implicit attitude
Punishment
Reward
Working memory

ABSTRACT

Experiments assessing the prevalence and magnitude of dishonesty have provided a large body of empirical findings regarding the cognitive nature of honesty. However, the personal factors that regulate dishonest behavior have yet to be fully clarified. This study examined two factors that potentially inhibit dishonesty—implicit attitudes toward dishonesty and executive control. In Study 1, the participants completed the Implicit Association Test (IAT), which measured their implicit attitudes toward dishonesty, and a working memory (WM) task, which was used to index executive control. The participants subsequently completed an incentivized coin-flip prediction task wherein they were given real and repeated opportunities for dishonest reward acquisition and punishment avoidance. The results revealed that individuals showing stronger negative implicit attitudes toward dishonesty engaged in a lower frequency of dishonest behavior for punishment avoidance, although this effect was marginal. In contrast, WM capacity was not associated with variations in dishonest reward acquisition and punishment avoidance. A follow-up experiment on other-serving dishonesty, where dishonest reward acquisition and punishment avoidance were credited to two other anonymous participants, revealed that neither implicit attitudes toward dishonesty nor WM capacity was associated with dishonest behavior. An additional preregistered experiment in Study 2 demonstrated that the association between implicit attitudes toward dishonesty and self-serving dishonesty for punishment avoidance was again marginal. While it is tempting to conclude that implicit attitudes toward dishonesty are associated with self-serving dishonesty, the present study provides only weak evidence that should be interpreted with great caution. Implications for the reliability of the IAT are discussed.

1. Introduction

What makes people honest or dishonest? Two competing hypotheses have been proposed (Greene & Paxton, 2009): The "will" hypothesis assumes that honesty requires executive control to suppress the temptation to cheat. In contrast, the "grace" hypothesis assumes that honesty flows automatically and without a need for executive control to suppress the temptation to cheat. While both hypotheses have received empirical support (e.g., Capraro, Schultz, & Rand, 2019; Mead, Baumeister, Gino, Schweitzer, & Ariely, 2009; Shalvi, Eldar, & Bereby-Meyer, 2012; van't Veer, Stel, & van Beest, 2014), a recent functional neuroimaging study suggested that a middle ground exists between these two hypotheses in terms of the prefrontal control system (Speer, Smidts, & Boksem, 2020). The study showed that the patterns of prefrontal activity responsible for controlled behavior differed in individuals who consistently behaved honestly and those who frequently cheated. Specifically, increased prefrontal activity was associated with a lower probability of cheating in individuals who frequently cheat, whereas it was associated with a higher probability of cheating in individuals who generally decide to be honest. Thus, the role of prefrontal control is thought to vary depending on an individual's "moral default", that is, their automatic disposition to behave honestly or dishonestly.

While the moral default hypothesis provides new insights into the cognitive nature of honesty, this idea has received little empirical support. Can the variability in moral default that is linked to honesty and dishonesty be measured using cognitive tasks? One potentially useful task is the Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998), which is widely used to measure the implicit attitudes

* Corresponding author at: Kokoro Research Center, Kyoto University, 46 Shimoadachi-cho, Yoshida Sakyo-ku, Kyoto 606-8501, Japan.
E-mail address: abe.nobuhito.7s@kyoto-u.ac.jp (N. Abe).

https://doi.org/10.1016/j.jesp.2022.104285
Received 23 February 2021; Received in revised form 21 December 2021; Accepted 10 January 2022
Available online 1 February 2022

of participants toward specific categories or concepts via their reaction times (RTs) while they are performing categorizations. Implicit attitudes measured by the IAT are believed to reflect automatic and often socially undesirable attitudes toward targets. Implicit attitudes regulate unfavorable behaviors through inhibition, disgust, or twinges of conscience (e.g., Lee, Ong, Parmar, & Amit, 2019; Ueda, Yanagisawa, Ashida, & Abe, 2017) and are often used to predict immoral judgments, such as managerial ethical decisions (Marquardt & Hoeger, 2009) and hiring discrimination (Agerström & Rooth, 2011). The implicit attitude toward dishonesty assessed by the IAT might reflect an individual's moral default that regulates self-serving dishonesty (Abe, 2020).

Notably, Jung and Lee (2009) conducted a study with the IAT, where participants who cheated to gain monetary compensation showed a greater implicit preference for deception than those who behaved honestly. However, two major limitations remained unaddressed. First, since Jung and Lee (2009) examined the involvement of implicit attitudes but not executive control, it is not yet known whether implicit attitudes are unique and independent predictors of dishonest behavior. Second, Jung and Lee (2009) examined self-serving dishonesty only; therefore, it is unclear whether the association between implicit attitudes and dishonest behavior is specific to self-serving dishonesty or can be generalized to other types of dishonesty.

The present study was designed to address these two issues by clarifying the joint contribution of two personal factors to the regulation of self-serving and other-serving dishonesty: implicit attitudes toward dishonesty and executive control. These two personal factors can be conceptualized within the framework of dual-process theory (e.g., Evans, 2008; Shiffrin & Schneider, 1977). Implicit attitudes toward dishonesty is associated with 'hot' automatic processes, and executive control is linked to 'cold' controlled processes. These two kinds of processes can sometimes separately influence decision-making processes, but at other times, interact with each other (e.g., Hofmann, Friese, & Wiers, 2008; Strack & Deutsch, 2004).

In the present study, we used single-category IAT (scIAT; Karpinski & Steinman, 2006) and a working memory (WM) task to measure implicit attitudes and executive control, respectively. In the scIAT, which is a modification of the standard IAT that measures the strength of evaluative associations with a single attitude object, we assessed individual differences in the implicit association between "dishonesty" and "pleasant". In the WM task, we assessed individual differences in WM capacity, which is known to be a valid measure of executive control and promotes deliberate processes of thought (Brewin & Beaton, 2002). For example, in the context of moral judgments, Moore, Clark, and Kane (2008) reported that people with greater WM capacity show more utilitarian responses in life-and-death moral dilemmas. This finding raises the possibility that the capacity of WM, which is linked to deliberate cognitive processes that exert willful control to regulate our automatic behaviors, also influences honest or dishonest decisions. Regarding the WM task, we used a computation span task in which the participants were asked to engage in a verification task and a recall task for math equations (Ackerman, Beier, & Boyle, 2002; Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000).

In addition to these two tasks, we asked participants to engage in an experiment assessing self-serving dishonesty using an incentivized coin-flip prediction task in which participants were given repeated opportunities to gain monetary rewards or avoid monetary punishment by lying about their accuracy in predicting random computerized coin flips (Gerlach, Teodorescu, & Hertwig, 2019; Shalvi & De Dreu, 2014). We also conducted a follow-up experiment on other-serving dishonesty where dishonest reward acquisition and punishment avoidance were credited to two other anonymous participants to determine the specificity of the effects of implicit attitudes and WM capacity on self-serving dishonesty. Inclusion of this self- and other-interest distinction was inspired by the previous literature in which people regarded other-serving dishonesty as more morally acceptable than self-serving dishonesty (e.g., Bussey, 1999; Gino, Ayal, & Ariely, 2013; Hayashi et al., 2014; Lindskold & Han, 1986; Lindskold & Walters, 1983). A neuroimaging study also suggested differences between the two types of dishonesty: compared with a self-serving goal, the altruistic goal of benefiting a charity was associated with reduced activity in the anterior insula, a region implicated in negative emotional states (Yin, Hu, Dynowski, Li, & Weber, 2017). These findings led to the hypothesis that the effects of moral default with regard to honesty or dishonesty on decision-making processes differed depending on the type of dishonesty.

For these two experiments assessing self-serving and other-serving dishonesty, we made the following predictions. First, based on Jung and Lee (2009) and previous findings on the differences in the psychological and neural processes underlying self-serving and other-serving dishonesty (e.g., Bussey, 1999; Gino et al., 2013; Hayashi et al., 2014; Lindskold & Han, 1986; Lindskold & Walters, 1983; Yin et al., 2017), we expected that the IAT scores would predict the frequency of self-serving dishonesty but not the frequency of other-serving dishonesty. Second, based on Speer et al. (2020), which showed the flexible nature of executive control, and several previous studies with the IAT (e.g., Hofmann et al., 2008; Klauer, Schmitz, Teige-Mocigemba, & Voss, 2010), we expected that there would be no linear relationship between WM capacity and dishonest behavior in either self-serving or other-serving dishonesty experiments. We also explored the possibility that there would be an interaction between implicit attitudes and WM capacity on dishonest behavior, such that WM capacity positively predicts dishonest behavior among people with negative implicit attitudes toward dishonesty and WM capacity negatively predicts dishonest behavior among people with positive implicit attitudes toward dishonesty.

2. Study 1

2.1. Methods

2.1.1. Participants

The results of the present experiment on self-serving dishonesty are based on data obtained from 68 participants (33 males and 35 females; age range: 20–39 years, mean = 26.8). A statistical power analysis was performed for sample size estimation by G*power 3.1.9.6 (Faul, Erdfelder, Buchner, & Lang, 2009). A sample size of 68 participants was required to reach a power of 0.8, with a medium effect size of $f^2 = 0.15$ (Cohen, 1988) for multiple regressions with two predictors (i.e., IAT and WM) and an α level of 0.05; data collection ceased when this number was reached. We also conducted a follow-up experiment assessing other-serving dishonesty with an additional 68 participants (33 males and 35 females; age range: 20–39 years, mean = 27.1) for which we modified the incentivized prediction task so that the participants could not earn any money for themselves; instead, their earnings were credited to two other anonymous participants. We therefore analyzed the data obtained from a total of 136 participants. All participants provided written informed consent to participate in this study in accordance with the protocol approved by the ethical committee of Kyoto University.

2.1.2. General procedures

The experiments were conducted on two separate days, one month apart (mean interval = 28.1 days, range: 28–34 days). On day 1, the participants completed the scIAT. They also completed questionnaires on their socioeconomic status (SES) for exploratory research (see Supplementary Information). On day 2, the participants returned to the laboratory and completed the WM task and coin-flip prediction task, with the latter presented as a task to measure the paranormal ability to predict the future. We introduced this interval between the IAT and the coin-flip prediction task with the purpose of preventing participants from engaging in the coin-flip prediction task while being explicitly aware of what was measured in the IAT. After the participants completed all tasks, they were informed of the true nature of the experiments. Although the participants were told that they could earn extra rewards depending on their performance (self-serving dishonesty experiment) or earn rewards for two other anonymous participants (other-serving dishonesty experiment) in the coin-flip prediction task, all participants received a fixed additional reward of 1000 Japanese yen (approximately 10 USD).

2.1.3. IAT

We used a modified version of the scIAT (Karpinski & Steinman, 2006) to measure the "dishonesty–pleasant" association and the "dishonesty–unpleasant" association (Fig. 1A). The tests were divided into two consecutive stages – "congruent" stage and "incongruent" stage – and all participants engaged in the task in this fixed order. There were 24 practice trials followed by 72 trials (24 trials × 3 blocks) in each stage.

The participants were asked to categorize each stimulus as quickly and accurately as possible by pressing the appropriate buttons in each phase, where pleasant and unpleasant pictures and words associated with dishonesty were randomly presented (see the Supplementary Information for details on the stimuli). In the congruent stage, the participants were asked to press the "F" key when they were presented with pleasant pictures and the "J" key when they were presented with unpleasant pictures or dishonest words. The proportion of dishonest words, pleasant pictures, and negative pictures was 7:10:7. In the incongruent stage, the participants were asked to press the "F" key when they were presented with positive pictures or dishonest words and the "J" key for negative pictures. The proportion of dishonest words, positive pictures, and negative pictures was 7:7:10. Category reminder words (i.e., "pleasant", "unpleasant", "dishonest") were presented at the top left and right of the computer screen, and the target stimuli were presented at the center of the screen. Each stimulus was presented for 1500 ms, and feedback on the accuracy of each response was also provided. If the participants did not respond within 1500 ms, a warning message ("Please respond in time!") appeared at the center of the screen for a duration of 500 ms. Each participant's IAT score (D) was calculated on the basis of their RTs in the test blocks by dividing the difference in the mean RT across blocks (congruent vs. incongruent) by the standard deviation of RTs of all trials (Greenwald, Nosek, & Banaji, 2003). Smaller D values indicated smaller differences in the RTs between the congruent and incongruent stages, reflecting less difficulty in responding to the association "dishonesty–pleasant".

2.1.4. WM task

To measure the participants' WM capacity, we used a computation span task in which the participants were asked to engage in a verification task and a recall task (Ackerman et al., 2002; Oberauer et al., 2000; Fig. 1B). In each trial, the participants were presented with multiple math equations for 6 s. They were asked to verify whether the presented equations were correct and remember the displayed solutions, irrespective of the accuracy. After the final equation of the trial was presented, the participants were asked to recall the solutions of the equations in the presented order. Each equation included two operations using digits from 1 to 10, and the provided solutions were always one-digit numbers. The set size ranged from three to ten equations/solutions with 3 trials for each set, resulting in a total of 24 trials. The equations were randomly selected from the stimulus pool used in Cantor and Engle (1993). The participants were familiarized with the task by completing practice trials. The WM score was calculated based on the accurate trial responses where the participant could recall the digits in the correct order, and these were weighted by set size for computing the final score (maximum score = 156). Here, we emphasize that it is difficult to measure the extent to which executive function is used during the decision phase of the coin-flip prediction task (see below), in which multiple cognitive processes are interacting with each other. This potentially complex nature of the coin-flip prediction task is one of the reasons we designed our experiment to examine WM capacity in an independent task.

2.1.5. Coin-flip prediction task

In our main experiment that assessed self-serving dishonesty, we used a coin-flip prediction task in which the participants had opportunities to gain a monetary reward or to avoid monetary punishment by lying about the accuracy of their predictions (e.g., Abe & Greene, 2014;

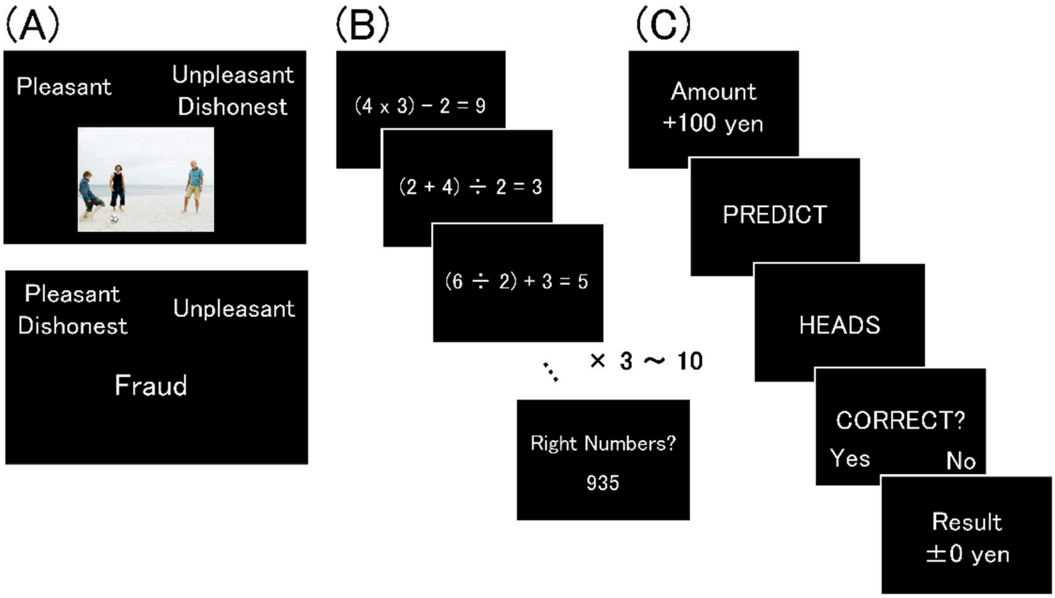

Fig. 1. Task sequence in the (A) IAT, (B) WM task, and (C) coin-flip prediction task. In the IAT (A), the participants were asked to categorize each stimulus as quickly and accurately as possible by pressing the appropriate buttons in the congruent and incongruent stages, where pleasant or unpleasant pictures and words associated with dishonesty were randomly presented. In the congruent stage (upper panel), 'dishonest' was linked to 'unpleasant', while in the incongruent stage, 'dishonest' was linked to 'pleasant' (lower panel). In the WM task (B), the participants were presented with multiple math equations and asked to verify whether the presented equations were correct and to remember the displayed solutions, irrespective of their accuracy. After the final equation of the trial was presented, the participants were asked to recall the solutions of the equations in the presented order. In the coin-flip task (C), the participants observed the trial's monetary value and privately predicted the outcome of the upcoming coin flip. The participant then observed the outcome of the coin flip and indicated whether the prediction was accurate. Then, the participant observed the amount of money won/lost based on the self-reported accuracy.

Abe, Greene, & Kiehl, 2018; Greene & Paxton, 2009; Hu, Pornpattananangkul, & Nusslock, 2015; Shalvi & De Dreu, 2014; Fig. 1C). We used a cover story to justify giving the participants obvious opportunities for dishonest gain. This task was presented as a measure of the paranormal ability to predict the future, aimed at testing the hypothesis that people are better able to predict the future when their predictions are (a) private and (b) financially incentivized. The participants were therefore implicitly led to believe that the opportunity for dishonest gain was a known but unintended byproduct of the experimental paradigm and that they were expected to behave honestly.

In each trial, the participants attempted to predict the outcomes of random computerized coin flips. The participants (1) were presented with the trial's monetary value (2 s), (2) privately predicted the outcome (heads or tails) of the upcoming coin flip (3 s), (3) were presented with the outcome of the coin flip (2 s), (4) indicated whether their prediction was accurate (3 s), (5) were presented with the amount of money won/lost (1 s), and (6) waited for the next trial (5 s). There were three different conditions (reward, punishment, and neutral) in the task, and each condition consisted of 10 trials. Therefore, the participants completed a total of 30 trials presented in a random order. In the reward condition, the participants could earn 100 Japanese yen (approximately 1 USD) for an accurate prediction of the coin flips in each trial, but there was no penalty for failing to predict the outcome of the coin flips. In the Punishment condition, the participants lost 100 Japanese yen for an accurate prediction of the coin flips in each trial; however, inaccurate predictions did not lead to a monetary loss. No money was at stake in the Neutral condition. Net losses were capped at 0 Japanese yen, and net winnings were capped at 1000 Japanese yen. However, all participants actually received 1000 yen as an additional reward. The instructions for the tasks were presented to the participants on a computer, as described in the Supplementary Information.

Using the coin-flip prediction task, we also conducted an other-serving dishonesty experiment as a follow-up study. Different individuals participated in the self-serving and other-serving dishonesty experiments, and no one participated in both experiments. We modified the incentive structure of the task such that the participants would not earn any money for themselves but could earn money that would be granted to two other anonymous participants. The participants were told that their rewards were determined by the performances of the two other participants (see the Supplementary Information for details). All other aspects of the experimental design and procedure were identical to the main experiment.

2.2. Results

2.2.1. Self-reported accuracy in the coin-flip prediction task

All of the statistical analyses were performed in the R programming environment (Version 4.0.5; R Core Team, 2021). For *t*-tests, we calculated the effect size as Cohen's d with Hedges's correction using the effsize package (Version 0.8.1; Torchiano, 2020) in R. In the self-serving dishonesty experiment, the mean proportions of self-reported accuracy were 52.6% (SD = 16.1%), 60.1% (SD = 20.3%), and 37.9% (SD = 21.2%) in the neutral, reward, and punishment conditions, respectively. To determine whether the participants behaved dishonestly in the reward and punishment conditions, one-way repeated analysis of variance (ANOVA) with condition as a factor was conducted. There was a significant main effect of condition on self-reported accuracy ($F(2, 134) = 21.86$, $p < .001$, $\eta_G^2 = 0.19$). Post hoc tests using Bonferroni's correction for multiple comparisons showed that the participants were less accurate in the punishment condition than in the neutral condition ($t(67) = 4.63$, $d = 0.77$, 95% confidence interval (CI) = [8.37, 21.04], adjusted $p < .001$) and reward condition ($t(67) = 5.40$, $d = 1.06$, 95% CI = [13.99, 30.42], adjusted $p < .001$). The difference between the reward condition and neutral condition was also significant ($t(67) = 2.65$, $d = 0.40$, 95% CI = [1.85, 13.15], adjusted $p = .030$).

In the other-serving dishonesty experiment, the mean proportions of self-reported accuracy were 51.0% (SD = 20.8%), 62.9% (SD = 20.0%), and 42.4% (SD = 19.3%) in the neutral, reward, and punishment conditions, respectively. One-way repeated measures ANOVA revealed a significant main effect of condition on self-reported accuracy ($F(2, 134) = 15.98$, $p < .001$, $\eta_G^2 = 0.15$). Post hoc tests using Bonferroni's correction for multiple comparisons revealed a significantly higher self-reported accuracy in the reward condition than in the neutral condition ($t(67) = 4.10$, $d = 0.58$, 95% CI = [6.11, 17.71], adjusted $p < .001$) and punishment condition ($t(67) = 4.86$, $d = 1.04$, 95% CI = [12.13, 29.04], adjusted $p < .001$). There was also a marginal difference in the self-reported accuracy in the punishment and neutral conditions ($t(67) = 2.34$, $d = 0.43$, 95% CI = [1.28, 16.07], adjusted $p = .067$).

2.2.2. Multiple regression analyses

We then conducted separate multiple linear regression analyses for each condition (i.e., reward and punishment) in the self-serving dishonesty experiment to predict the self-reported accuracy based on IAT D-scores (see Supplementary Information), WM scores, and their interaction. We also included sex (dummy-coded before being entered into the models; female = 0, male = 1) and age as control variables. All variables were standardized with a mean of 0 and standard deviation of 1 prior to the analyses. Zero-order correlations among the variables were exploratorily calculated and are summarized in Tables S1 and S2. In the regression model for the reward condition, no significant effect of IAT scores (standardized coefficient = −0.12, $t(62) = -0.92$, 95% CI = [−0.38, 0.14], $p = .36$), WM scores (standardized coefficient = −0.06, $t(62) = -0.41$, 95% CI = [−0.38, 0.25], $p = .69$), or their interaction (standardized coefficient = 0.03, $t(62) = 0.18$, 95% CI = [−0.26, 0.32], $p = .86$) was observed (Table 1). The model fit was not statistically significant ($F(5,62) = 0.67$, $p = .64$, $R^2 = 0.05$). In contrast, we found that the IAT score was a marginally significant predictor of self-reported accuracy in the punishment condition (standardized coefficient = 0.23, $t(62) = 1.85$, 95% CI = [−0.02, 0.47], $p = .069$; Table 2). The effects of WM scores and the interaction between IAT and WM scores were not significant (WM: standardized coefficient = −0.24, $t(62) = -1.64$, 95% CI = [−0.53, 0.05], $p = .11$; interaction: standardized coefficient = 0.23, $t(62) = 1.70$, 95% CI = [−0.04, 0.50], $p = .093$). The model fit was statistically significant ($F(5,62) = 2.57$, $p = .035$, $R^2 = 0.17$).

Next, we tested whether the IAT scores and WM scores predicted other-serving dishonesty using the same multiple regression analyses employed in the reward and punishment conditions (Tables 3 and 4, respectively). No significant effects of IAT scores were observed on other-serving dishonesty for reward earnings (standardized coefficient = 0.0002, $t(62) = 0.002$, 95% CI = [−0.25, 0.26], $p = 1.00$) or punishment avoidance (standardized coefficient = −0.02, $t(62) = -0.14$, 95% CI = [−0.28, 0.24], $p = .89$). We also observed no effects of WM scores (reward condition: standardized coefficient = 0.13, $t(62) = 1.01$, 95% CI = [−0.13, 0.40], $p = .32$; punishment condition: standardized coefficient = 0.02, $t(62) = 0.15$, 95% CI = [−0.25, 0.29], $p = .88$) or

Table 1
Results of the multiple regression analysis predicting self-serving dishonesty in the reward condition.

	Standardized coefficient	Standard error	*t*	*p*	95% CI for the standardized coefficient	
(Intercept)	0.01	0.13	0.06	0.96	−0.25	0.26
IAT D-score	−0.12	0.13	−0.92	0.36	−0.38	0.14
WM score	−0.06	0.16	−0.41	0.69	−0.38	0.25
IAT D-score × WM score	0.03	0.14	0.18	0.86	−0.26	0.32
age	−0.10	0.14	−0.73	0.47	−0.37	0.17
sex	−0.17	0.13	−1.33	0.19	−0.42	0.08
$R^2 = 0.05$						

IAT, Implicit Association Test; WM, working memory; sex was dummy-coded before being entered into the models (female = 0, male = 1).

Table 2
Results of the multiple regression analysis predicting self-serving dishonesty in the punishment condition.

	Standardized coefficient	Standard error	*t*	*p*	95% CI for the standardized coefficient	
(Intercept)	0.06	0.12	0.51	0.61	−0.18	0.30
IAT D-score	0.23	0.12	1.85	0.069	-0.02	0.47
WM score	−0.24	0.15	−1.64	0.11	−0.53	0.05
IAT D-score × WM score	0.23	0.14	1.70	0.093	−0.04	0.50
age	0.16	0.13	1.22	0.23	−0.10	0.41
sex	−0.08	0.12	−0.69	0.50	−0.32	0.15
$R^2 = 0.17$						

IAT, Implicit Association Test; WM, working memory; sex was dummy-coded before being entered into the models (female = 0, male = 1).

Table 3
Results of the multiple regression analysis predicting other-serving dishonesty in the reward condition.

	Standardized coefficient	Standard error	*t*	*p*	95% CI for the standardized coefficient	
(Intercept)	0.01	0.13	0.09	0.93	−0.24	0.26
IAT D-score	0.0002	0.13	0.002	1.00	−0.25	0.26
WM score	0.13	0.13	1.01	0.32	−0.13	0.40
IAT D-score × WM score	−0.08	0.12	−0.64	0.53	−0.32	0.16
age	0.09	0.13	0.68	0.50	−0.17	0.35
sex	−0.07	0.13	−0.59	0.56	−0.33	0.18
$R^2 = 0.03$						

IAT, Implicit Association Test; WM, working memory; sex was dummy-coded before being entered into the models (female = 0, male = 1).

Table 4
Results of the multiple regression analysis predicting other-serving dishonesty in the punishment condition.

	Standardized coefficient	Standard error	*t*	*p*	95% CI for the standardized coefficient	
(Intercept)	0.007	0.13	0.06	0.95	−0.25	0.26
IAT D-score	−0.02	0.13	−0.14	0.89	−0.28	0.24
WM score	0.02	0.13	0.15	0.88	−0.25	0.29
IAT D-score × WM score	−0.05	0.12	−0.44	0.67	−0.30	0.19
age	0.04	0.13	0.29	0.77	−0.23	0.30
sex	0.07	0.13	0.55	0.58	−0.18	0.33
$R^2 = 0.01$						

IAT, Implicit Association Test; WM, working memory; sex was dummy-coded before being entered into the models (female = 0, male = 1).

Table 5
Results of the multiple regression analysis predicting dishonesty in the punishment condition based on group and IAT score.

	Standardized coefficient	Standard error	*t*	*p*	95% CI for the standardized coefficient	
(Intercept)	−0.02	0.09	−0.28	0.78	−0.19	0.15
IAT D-score	0.13	0.09	1.44	0.15	−0.05	0.30
group	0.13	0.09	1.46	0.15	−0.04	0.30
IAT D-score × group	−0.15	0.09	−1.78	0.078	−0.33	0.02
age	0.11	0.09	1.23	0.22	−0.06	0.28
sex	−0.02	0.09	−0.27	0.78	−0.19	0.15
$R^2 = 0.07$						

IAT, Implicit Association Test; WM, working memory; sex (female = 0, male = 1) and group (self-serving = 0, other-serving = 1) were dummy-coded before being entered into the models.

interaction between IAT and WM (reward condition: standardized coefficient = −0.08, $t(62) = -0.64$, 95% CI = [−0.32, 0.16], $p = .53$; punishment condition: standardized coefficient = −0.05, $t(62) = -0.44$, 95% CI = [−0.30, 0.19], $p = .67$). Moreover, the model fit was not significant for either the reward condition model ($F(5, 62) = 0.42$, $p = .83$, $R^2 = 0.03$) or the punishment condition model ($F(5, 62) = 0.11$, $p = 0.99$, $R^2 = 0.01$). In short, IAT scores, WM capacity, and their interaction were not significant predictors of other-serving dishonesty.

Here, the critical test was to determine whether the prediction using IAT scores for self-reported accuracy in the punishment condition was specific to self-serving dishonesty (Table 5). We performed a multiple linear regression to predict the self-reported accuracy in the punishment condition based on the group (i.e., self-serving vs. other-serving dishonesty; dummy-coded before being entered into the models; self-serving = 0, other-serving = 1) and IAT scores while controlling for sex and age. As we observed no significant involvement of WM scores in the prediction of dishonesty in separate regression analyses (see above), we did not include WM scores in this model. The interaction between IAT scores and group was marginally significant (standardized coefficient = −0.15, $t(130) = -1.78$, 95% CI = [−0.33, 0.02], $p = .078$; Fig. 2). The model fit was not significant ($F(5,130) = 1.89$, $p = .10$, $R^2 = 0.07$). These results partly confirmed the contribution of an implicit attitude toward dishonesty to self-serving dishonesty, but not other-serving dishonesty, for punishment avoidance.

3. Study 2

In Study 1, the following two findings were obtained and needed replication. First, the participants showing stronger negative implicit attitudes toward dishonesty engaged in a lower frequency of selfish dishonest behavior for punishment avoidance (though marginal) but not for reward acquisition. While the results of the association between implicit attitudes and dishonest behavior were consistent with our a priori prediction, the contrast between punishment avoidance and reward acquisition was an unexpected result. Second, the interaction between implicit attitudes toward dishonesty and WM capacity was marginal, indicating that a firm conclusion could not be drawn.

We therefore conducted an additional experiment with

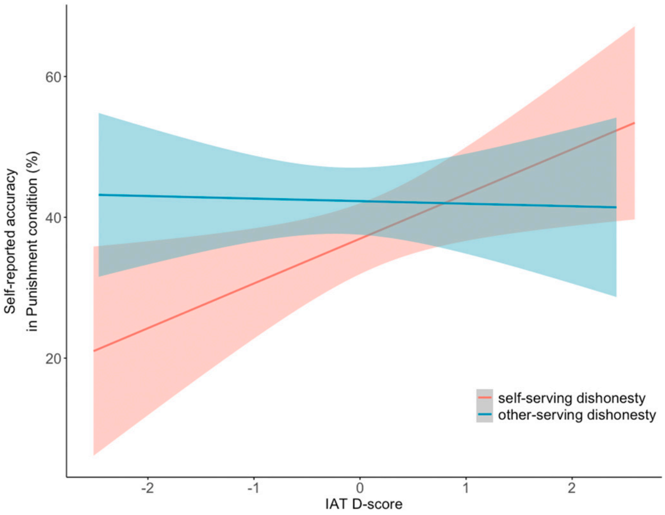

Fig. 2. Results of the regression analysis predicting self-reported accuracy in the punishment condition of the coin-flip prediction task based on the IAT D-score, the group (i.e., self-serving or other-serving dishonesty experiment), and their interaction. The shaded areas represent 95% confidence intervals.

preregistration to determine (a) whether the effects of implicit attitudes toward dishonesty on dishonest behavior were specific to the context of punishment avoidance and (b) whether there was an interaction between implicit attitudes toward dishonesty and WM capacity. Our preregistration details can be found at Aspredicted.org (https://aspredicted.org/f6xn2.pdf).

3.1. Methods

3.1.1. Participants

We determined the sample size based on the effect size observed in Study 1. A-priori power analysis using a G* Power (Version: 3.1.9.6; Faul et al., 2009) for the multiple regression with 3 predictors (IAT score, WM score, and their interaction; "Linear multiple regression: Fixed model, R^2 deviation from zero") suggested that to achieve 80% power to detect the effect size of Study 1, 54 participants were needed (see the Supplementary Information). Assuming an exclusion rate of approximately 10% (e.g., no shows and cancellations), it was determined that a sample size of 60 was required. We therefore recruited 60 participants (30 males and 30 females) aged 20–39 years who had not participated in Study 1. The data from two participants were excluded from all analyses: one due to the cancellation of the second part of the experiment (day 2) and the other due to technical errors with the IAT data acquisition. Consequently, our final sample size was 58 (29 males and 29 females; age range: 20–38 years, mean = 24.5). Written informed consent was provided by all participants prior to their study participation in accordance with the protocol accepted by the ethical committee of Kyoto University.

3.1.2. Procedures

All tasks and procedures were identical to the self-serving dishonesty experiment in Study 1 with the exception of the following two points. First, as our primary hypothesis concerned self-serving dishonesty, we conducted the replication study only for self-serving dishonesty. Second, we established a 28-day interval between the WM task and coin-flip task to remove possible effects of WM task performance on participant performance in the coin-flip task. That is, the participants completed the IAT and the WM task on day 1 and then engaged in the coin-flip prediction task on day 2.

3.2. Results

For the neutral, reward, and punishment conditions in the coin-flip prediction task, the mean proportions of self-reported accuracy were 51.4% (SD = 15.6%), 68.3% (SD = 20.7%), and 35.0% (SD = 19.8%), respectively. Similar to Study 1, ANOVA revealed that participants behaved dishonestly in both the reward and punishment conditions. There was a significant main effect of condition on self-reported accuracy ($F(2, 114) = 43.63$, $p < .001$, $\eta_G{}^2 = 0.35$). Post hoc tests using Bonferroni's correction for multiple comparisons showed that the participants were less accurate in the punishment condition than in the neutral condition ($t(57) = 5.37$, $d = 0.91$, 95% CI = [10.27, 22.49], adjusted $p < .001$) and reward condition ($t(57) = 7.51$, $d = 1.62$, 95% CI = [24.41, 42.15], adjusted $p < .001$). They were also more accurate in the reward condition than in the neutral condition ($t(57) = 5.59$, $d = 0.90$, 95% CI = [10.84, 22.95], adjusted $p < .001$).

We then conducted preregistered separate multiple linear regression analyses for each condition (i.e., reward and punishment conditions) to predict self-reported accuracy based on IAT D-scores, WM scores and their interaction, with sex (dummy-coded before being entered into the models; female = 0, male = 1) and age as control variables. All variables were standardized with a mean of 0 and standard deviation of 1 prior to the analyses. In the regression model for the reward condition, although unexpected, WM scores were a significant predictor of self-reported accuracy (standardized coefficient = −0.36, $t(52) = -2.37$, 95% CI = [−0.67, −0.06], $p = .022$). However, no significant effect of IAT scores (standardized coefficient = −0.19, $t(52) = -1.42$, 95% CI = [−0.46, 0.08], $p = .16$) or their interaction (standardized coefficient = −0.07, $t(52) = -0.51$, 95% CI = [−0.33, 0.20], $p = .61$) was observed (Table 6). The model fit was not statistically significant ($F(5,52) = 1.67$, $p = .16$, $R^2 = 0.14$).

On the other hand, IAT scores were a marginally significant predictor of self-reported accuracy in the punishment condition (standardized coefficient = 0.24, $t(52) = 1.75$, 95% CI = [−0.04, 0.52], $p = .086$; Table 7). The effects of WM scores and the interaction between IAT and WM scores were not significant (WM: standardized coefficient = −0.04, $t(52) = -0.23$, 95% CI = [−0.35, 0.28], $p = .82$; interaction: standardized coefficient = −0.16, $t(52) = -1.16$, 95% CI = [−0.43, 0.11], $p = .25$). The model fit was not statistically significant ($F(5,52) = 1.10$, $p = .37$, $R^2 = 0.10$). In summary, this preregistered experiment demonstrated that IAT scores were a marginally significant predictor of dishonest punishment avoidance and that there were no interaction effects between implicit attitudes and WM capacity.

4. Discussion

There has been a longstanding debate in psychology about what makes humans honest or dishonest. We hypothesized that an implicit attitude toward dishonesty is an important determinant of self-serving dishonest behavior. The results obtained in Study 1 were partly consistent with this hypothesis. That is, implicit attitudes toward dishonesty were marginally associated with the frequency of self-serving dishonesty to avoid monetary punishment but not to acquire a monetary reward, and WM capacity was not a significant predictor of self-serving dishonesty. An interaction between implicit attitudes and WM capacity did not reach significance. A follow-up experiment revealed that neither an implicit attitude nor WM capacity was a predictor of other-serving dishonesty to gain a reward or avoid punishment, indicating the specificity of the contribution of implicit attitudes to self-serving dishonesty. In Study 2, which was designed as a replication of the results from the self-serving dishonesty experiment, the effects of implicit attitudes on dishonesty to avoid monetary punishment were again marginal. In light of these observations, we first discuss issues regarding the reliability of the IAT and then discuss the present findings with attention to the limitations.

The IAT has been criticized for a long time, especially recently (e.g., Arkes & Tetlock, 2004; Mitchell & Tetlock, 2017). One of the hotly debated topics that is relevant to the present study is the reliability of the IAT. For example, Gawronski et al. (2017) recently reported that implicit measures for the domains of self-concept, racial attitudes, and political attitudes, assessed by the IAT and affect misattribution procedure, showed significantly lower stability over time (weighted average $r = 0.54$) than their corresponding explicit measures (weighted average $r = 0.75$). Given this modest within-subject reliability, it was not surprising that the effects of the IAT were only marginal in both Study 1 and Study 2 (see also Tello, Harika-Germaneau, Serra, Jaafari, &

Table 6
Results of the multiple regression analysis predicting self-serving dishonesty in the reward condition.

	Standardized coefficient	Standard error	t	p	95% CI for the standardized coefficient	
(Intercept)	−0.01	0.13	−0.10	0.93	−0.27	0.25
IAT D-score	−0.19	0.13	−1.42	0.16	−0.46	0.08
WM score	−0.36	0.15	−2.37	0.022	−0.67	−0.06
IAT D-score × WM score	−0.07	0.13	−0.51	0.61	−0.33	0.20
age	0.03	0.15	0.18	0.86	−0.27	0.32
sex	0.27	0.15	1.81	0.076	−0.03	0.57
$R^2 = 0.14$						

IAT, Implicit Association Test; WM, working memory; sex was dummy-coded before being entered into the models (female = 0, male = 1).

Table 7
Results of the multiple regression analysis predicting self-serving dishonesty in the punishment condition.

	Standardized coefficient	Standard error	*t*	*p*	95% CI for the standardized coefficient	
(Intercept)	−0.03	0.13	−0.21	0.83	−0.30	0.24
IAT D-score	0.24	0.14	1.75	0.086	−0.04	0.52
WM score	−0.04	0.16	−0.23	0.82	−0.35	0.28
IAT D-score × WM score	−0.16	0.14	−1.16	0.25	−0.43	0.11
age	−0.18	0.15	−1.20	0.24	−0.48	0.12
sex	−0.13	0.15	−0.86	0.39	−0.44	0.17
$R^2 = 0.10$						

IAT, Implicit Association Test; WM, working memory; sex was dummy-coded before being entered into the models (female = 0, male = 1).

Chatard, 2020). However, we believe that the present results do not necessarily invalidate our conclusions. That is, the overall pattern of the results obtained from Study 1 and Study 2 was still similar in terms of the possible association between implicit attitudes toward dishonesty and dishonest behavior to avoid monetary punishment. In addition, the fact that these results are consistent with a priori prediction and the previous literature (Jung & Lee, 2009) reduces the likelihood that the present results were obtained by chance. Further support comes from a previous study with a large sample showing that the IAT (including the scIAT used in the present study) showed good psychometric qualities with superior discriminant validity, fair reliability, and convergent validity (Bar-Anan & Nosek, 2014). Note, however, that we do not wish to imply that recent criticisms of the IAT are trivial. Instead, we regard the IAT as an only modestly useful measure of implicit attitudes with limited, but at least acceptable, reliability that requires careful interpretation of the results.

With these caveats in mind, we now turn to the discussion of the association between implicit attitudes and dishonest behavior. Our finding that the results regarding the IAT were marginal only in the punishment condition and not in the reward condition is likely to reflect that the punishment condition better captures individual differences in self-serving dishonesty that might be derived from the greater motivation to be dishonest in a context of punishment avoidance than reward acquisition. People are willing to exert more effort to avoid a loss than to obtain a gain of a similar size (Tversky & Kahneman, 1991). This phenomenon, known as loss aversion, has been observed in many psychological and behavioral economics studies (e.g., Neumann & Böckenholt, 2014; Ruggeri et al., 2020). Notably, this tendency substantially influences ethical judgments and moral decision-making, including those leading to dishonest behavior (e.g., Kern & Chugh, 2009; Grolleau, Kocher, & Sutan, 2016; Schindler & Pfattheicher, 2017, but see Soraperra, Weisel, & Ploner, 2019). For example, Kern and Chugh (2009) reported that participants in the loss-frame condition were more likely to favor gathering “insider information” and lying more than participants in the gain-frame condition. Consistent with these previous findings, the present study consistently revealed that the self-reported accuracy in the punishment condition was significantly lower than that in the neutral condition in the self-serving dishonesty experiments across the two studies.

The present results on the association between WM capacity and the likelihood of dishonest behavior were mixed; while there were no significant associations between WM capacity and dishonesty in any condition in Study 1, WM scores were a significant (but unexpected) predictor of self-reported accuracy in the reward condition in Study 2. We do not know the precise reason for these divergent effects, but overall, our findings did not support the more intuitive, simple hypothesis that a person with greater executive control is more consistently honest (Bereby-Meyer & Shalvi, 2015). Instead, executive control is likely to play a flexible role in overriding the automatic disposition to behave honestly or dishonestly based on an individual’s moral default (Speer et al., 2020). The present findings highlight the possibility that while the components of automatic systems, including the implicit attitudes measured here and reward sensitivity (Abe & Greene, 2014), are important determinants of honesty or dishonesty, the components of deliberate systems, including executive control, have nonlinear or limited effects on modulating honest or dishonest behavior.

As we hypothesized, an implicit attitude toward dishonesty was not a predictor of other-serving dishonesty, regardless of whether it was for reward acquisition or punishment avoidance. Given that other-serving dishonesty is often regarded as socially or morally acceptable (e.g., Bussey, 1999; Gino et al., 2013; Hayashi et al., 2014; Lindskold & Han, 1986; Lindskold & Walters, 1983), it might not represent the typical “dishonesty” of participants at the conceptual level. Consistent with this idea, researchers have proposed that lying is not a homogeneous concept and that not all types of lies are automatically considered to represent negative values in terms of moral norms (e.g., Talwar & Lee, 2012; Wu, Loke, Xu, & Lee, 2011). While we do not have the data to directly test this possibility, our findings indicated that, at least in the present experimental paradigm, an implicit attitude toward dishonesty does not influence other-serving dishonesty. Likewise, the contribution of WM capacity to other-serving dishonesty might be relatively limited. Other-serving dishonesty can be more easily justified due to reduced moral conflict (Gino et al., 2013), which might require the engagement of executive control to a lesser extent.

This study has several limitations. First, we note once again that the effects of implicit attitudes toward dishonesty on self-serving dishonesty to avoid monetary punishment were only marginal across the two studies. The fact that this effect was consistent with a priori hypotheses and previous findings (Jung & Lee, 2009) reduces the likelihood that the effects were due to chance, although these results should be interpreted with great caution. Second, since we assessed WM capacity of the participants in an independent task, it did not capture the amount of executive control the participants actually exerted while making their decisions in the coin-flip prediction task. Third, we cannot assert that the WM task is the most appropriate proxy for executive control: the Stroop task or go/no-go task might be more appropriate to measure individual differences in executive control. Despite these limitations, the present study provided weak but novel evidence that implicit attitudes toward dishonesty are associated with self-serving dishonesty to avoid monetary punishment. We speculate that the implicit attitude measured here is closely linked to an individual’s moral default for self-serving dishonesty.

Data availability

The data are available from the corresponding author upon reasonable request.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

We are grateful to Aoi Naito, Daichi Yokohata, Koichiro Inada, and Tatsuyoshi Ogawa for their assistance with data collection. This research was supported by JSPS KAKENHI Grant Numbers 16H05861 and 19H00628.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2022.104285.

References

Abe, N. (2020). Overriding a moral default for honesty or dishonesty. *Proceedings of the National Academy of Sciences of the United States of America, 117*(36), 21844–21846. https://doi.org/10.1073/pnas.2014489117

Abe, N., & Greene, J. D. (2014). Response to anticipated reward in the nucleus accumbens predicts behavior in an independent test of honesty. *Journal of Neuroscience, 34*(32), 10564–10572. https://doi.org/10.1523/JNEUROSCI.0217-14.2014

Abe, N., Greene, J. D., & Kiehl, K. A. (2018). Reduced engagement of the anterior cingulate cortex in the dishonest decision-making of incarcerated psychopaths. *Social Cognitive and Affective Neuroscience, 13*(8), 797–807.

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General, 131*(4), 567–589. https://doi.org/10.1037/0096-3445.131.4.567

Agerström, J., & Rooth, D.-O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *The Journal of Applied Psychology, 96*(4), 790–805. https://doi.org/10.1037/a0021594

Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'Fail' the implicit association test?". *Psychological Inquiry, 15*(4), 257–278. https://doi.org/10.1207/s15327965pli1504_01

Bar-Anan, Y., & Nosek, B. A. (2014). A comparative investigation of seven indirect attitude measures. *Behavior Research Methods, 46*(3), 668–688. https://doi.org/10.3758/s13428-013-0410-6

Bereby-Meyer, Y., & Shalvi, S. (2015). Deliberate honesty. *Current Opinion in Psychology, 6*, 195–198. https://doi.org/10.1016/j.copsyc.2015.09.004

Brewin, C. R., & Beaton, A. (2002). Thought suppression, intelligence, and working memory capacity. *Behaviour Research and Therapy, 40*(8), 923–930. https://doi.org/10.1016/S0005-7967(01)00127-9

Bussey, K. (1999). Children's categorization and evaluation of different types of lies and truths. *Child Development, 70*(6), 1338–1347. https://doi.org/10.1111/1467-8624.00098

Cantor, J., & Engle, R. W. (1993). Working-memory capacity as long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(5), 1101–1114. https://doi.org/10.1037/0278-7393.19.5.1101

Capraro, V., Schultz, J., & Rand, D. G. (2019). Time pressure and honesty in a deception game. *Journal of Behavioral and Experimental Economics, 79*, 93–99. https://doi.org/10.1016/j.socec.2019.01.007

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43*(3), 300–312. https://doi.org/10.1177/0146167216684131

Gerlach, P., Teodorescu, K., & Hertwig, R. (2019). The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin, 145*(1), 1–44. https://doi.org/10.1037/bul0000174

Gino, F., Ayal, S., & Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior & Organization, 93*, 285–292.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences of the United States of America, 106*(30), 12506–12511. https://doi.org/10.1016/j.jebo.2013.04.005

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037//0022-3514.74.6.1464

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197

Grolleau, G., Kocher, M. G., & Sutan, A. (2016). Cheating and loss aversion: Do people cheat more to avoid a loss? *Management Science, 62*(12), 3428–3438. https://doi.org/10.1287/mnsc.2015.2313

Hayashi, A., Abe, N., Fujii, T., Ito, A., Ueno, A., Koseki, Y., Mugikura, S., Takahashi, S., & Mori, E. (2014). Dissociable neural systems for moral judgment of anti- and pro-social lying. *Brain Research, 1556*, 46–56. https://doi.org/10.1016/j.brainres.2014.02.011

Hofmann, W., Friese, M., & Wiers, R. W. (2008). Impulsive versus reflective influences on health behavior: A theoretical framework and empirical review. *Health Psychology Review, 2*(2), 111–137. https://doi.org/10.1080/17437190802617668

Hu, X., Pornpattananangkul, N., & Nusslock, R. (2015). Executive control- and reward-related neural processes associated with the opportunity to engage in voluntary dishonest moral decision making. *Cognitive, Affective, & Behavioral Neuroscience, 15*(2), 475–491. https://doi.org/10.3758/s13415-015-0336-9

Jung, K. H., & Lee, J. H. (2009). Implicit and explicit attitude dissociation in spontaneous deceptive behavior. *Acta Psychologica, 132*(1), 62–67. https://doi.org/10.1016/j.actpsy.2009.06.004

Karpinski, A., & Steinman, R. B. (2006). The single category implicit association test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91*(1), 16–32. https://doi.org/10.1037/0022-3514.91.1.16

Kern, M. C., & Chugh, D. (2009). Bounded ethicality: The perils of loss framing. *Psychological Science, 20*(3), 378–384. https://doi.org/10.1111/j.1467-9280.2009.02296.x

Klauer, K. C., Schmitz, F., Teige-Mocigemba, S., & Voss, A. (2010). Understanding the role of executive control in the implicit association test: Why flexible people have small IAT effects. *Quarterly Journal of Experimental Psychology, 63*(3), 595–619. https://doi.org/10.1080/17470210903076826

Lee, J. J., Ong, M., Parmar, B., & Amit, E. (2019). Lay theories of effortful honesty: Does the honesty-effort association justify making a dishonest decision? *Journal of Applied Psychology, 104*(5), 659–677. https://doi.org/10.1037/apl0000364

Lindskold, S., & Han, G. (1986). Intent and the judgment of lies. *Journal of Social Psychology, 126*, 129–130. https://doi.org/10.1080/00224545.1986.9713581

Lindskold, S., & Walters, P. S. (1983). Categories for acceptability of lies. *Journal of Social Psychology, 120*, 129–136. https://doi.org/10.1080/00224545.1983.9712018

Marquardt, N., & Hoeger, R. (2009). The effect of implicit moral attitudes on managerial decision-making: An implicit social cognition approach. *Journal of Business Ethics, 85*(2), 157–171. https://doi.org/10.1007/s10551-008-9754-8

Mead, N. L., Baumeister, R. F., Gino, F., Schweitzer, M. E., & Ariely, D. (2009). Too tired to tell the truth: Self-control resource depletion and dishonesty. *Journal of Experimental Social Psychology, 45*(3), 594–597. https://doi.org/10.1016/j.jesp.2009.02.004

Mitchell, G., & Tetlock, P. E. (2017). Popularity as a poor proxy for utility: The case of implicit prejudice. In S. O. Lilienfeld, & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions (pp. 164–195)*. https://doi.org/10.1002/9781119095910.ch10. Wiley Blackwell.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*(6), 549–557. https://doi.org/10.1111/j.1467-9280.2008.02122.x

Neumann, N., & Böckenholt, U. (2014). A meta-analysis of loss aversion in product choice. *Journal of Retailing, 90*(2), 182–197. https://doi.org/10.1016/j.jretai.2014.02.002

Oberauer, K., Süß, H.-M., Schulze, R. R., Wilhelm, O. O., & Wittmann, W. W. (2000). Working memory capacity—Facets of a cognitive ability construct. *Personality and Individual Differences, 29*(6), 1017–1045. https://doi.org/10.1016/S0191-8869(99)00251-2

R Core Team. (2021). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., … Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour, 4*(6), 622–633.

Schindler, S., & Pfattheicher, S. (2017). The frame of the game: Loss-framing increases dishonest behavior. *Journal of Experimental Social Psychology, 69*, 172–177. https://doi.org/10.1016/j.jesp.2016.09.009

Shalvi, S., & De Dreu, C. K. (2014). Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences of the United States of America, 111*(15), 5503–5507. https://doi.org/10.1073/pnas.1400724111

Shalvi, S., Eldar, O., & Bereby-Meyer, Y. (2012). Honesty requires time (and lack of justifications). *Psychological Science, 23*(10), 1264–1270. https://doi.org/10.1177/0956797612443835

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review, 84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Soraperra, I., Weisel, O., & Ploner, M. (2019). Is the victim max (Planck) or Moritz? How victim type and social value orientation affect dishonest behavior. *Journal of Behavioral Decision Making, 32*(2), 168–178. https://doi.org/10.1002/bdm.2104

Speer, S. P. H., Smidts, A., & Boksem, M. A. S. (2020). Cognitive control increases honesty in cheaters, but cheating in those who are honest. *Proceedings of the National Academy of Sciences of the United States of America, 117*(32), 19080–19091. https://doi.org/10.1073/pnas.2003480117

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

van't Veer, A. E., Stel, M., & van Beest, I. (2014). Limited capacity to lie: Cognitive load interfere with being dishonest. *Judgment and Decision making, 9*(3), 199–206. https://doi.org/10.2139/ssrn.2351377

Talwar, V., & Lee, K. (2012). Little liars: Origins of verbal deception in children. In K. Fujita, & S. Itakura (Eds.), *Origins of the social mind: Evolutionary and developmental views* (pp. 157–178). Springer Japan: Tokyo.

Tello, N., Harika-Germaneau, G., Serra, W., Jaafari, N., & Chatard, A. (2020). Forecasting a fatal decision: Direct replication of the predictive validity of the suicide-implicit association test. *Psychological Science, 31*(1), 65–74. https://doi.org/10.1177/0956797619893062

Torchiano, M. (2020). *Efficient effect size computation. R package version 0.8.1.* https://doi.org/10.5281/zenodo.1480624

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *Quaterly Journal of Economics, 106*(4), 1039–1061. https://doi.org/10.2307/2937956

Ueda, R., Yanagisawa, K., Ashida, H., & Abe, N. (2017). Implicit attitudes and executive control interact to regulate interest in extra-pair relationships. *Cognitive, Affective, &*

Behavioral Neuroscience, 17(6), 1210–1220. https://doi.org/10.3758/s13415-017-0543-7

Wu, D., Loke, I. C., Xu, F., & Lee, K. (2011). Neural correlates of evaluations of lying and truth-telling in different social contexts. *Brain Research, 1389*, 115–124. https://doi.org/10.1016/j.brainres.2011.02.084

Yin, L., Hu, Y., Dynowski, D., Li, J., & Weber, B. (2017). The good lies: Altruistic goals modulate processing of deception in the anterior insula. *Human Brain Mapping, 38*(7), 3675–3690. https://doi.org/10.1002/hbm.23623