Honesty pledges: The effects of involvement and identification over time

Running head: Honesty Pledges

Eyal Peer[1], Nina Mazar[2], Yuval Feldman[3], Dan Ariely[4]

[1] School of Public Policy and Governance, Hebrew University of Jerusalem

[2] Questrom School of Business, Boston University

[3] Faculty of Law, Bar-Ilan University

[4] Fuqua School of Business, Duke University

Corresponding address: Eyal Peer, eyal.peer@mail.huji.ac.il, Room 1742, Mexico Wing, Mount Scopus, 59000, Jerusalem, Israel. + 972-544-955460.

Honesty pledges: The effects of involvement and identification over time

Abstract

Authorities and managers often rely on individuals' and businesses' self-reports, and employ various kinds of veracity statements, honesty pledges or oaths to ensure they do not over-claim payments, benefits, or other resources. While some research show honesty pledges can reduce dishonesty, others have provided mixed, and sometimes even contradictory, findings. We argue that understanding and predicting when honesty pledges are effective has been obstructed due to variations in the operationalizations of honesty pledges in previous research. Specifically, we identify that previous studies varied in whether and how the pledge asks individuals to identify (e.g., by ID, signature or name) and how much involvement, if at all, the pledge requires from the individual. In four pre-registered studies (N>5,000), we systematically examine these elements of a pledge to find that increasing involvement of pledgers (by having them copy the text of the pledge) is often more effective than those that only require identification. In contrast, pledges that only require individuals to read and agree are mostly ineffective. Moreover, we find that the effects of high-involvement pledges persist both over time and both after a short delay between the pledge and the opportunity to cheat. Together, these results contribute both theoretically to the understanding of the mechanisms underlying different honesty pledges as well as offer practical advice to managers and policymakers on how to effectively prevent or reduce dishonesty in self-reports.

Honesty pledges: The effects of involvement and identification over time

In their attempts to reduce or prevent unethical behaviors, or to ensure compliance with policies and regulations, authorities often ask individuals or businesses to declare that they are or will be reporting or acting in abidance with relevant rules, regulations, and/or standards. For example, witnesses are asked to swear to tell the truth in their testimony, students must state they will not cheat on their exam, employees need to sign contracts with honesty clauses, vendors are required to submit that their products and procedures follow regulations, and business owners sign statements when applying for permits or benefits. Although such declarations should encourage ethical behavior (de Bruin, 2016; Schlesinger, 2011), when they are relied upon instead of monitoring and auditing, they also provide an opportunity for false, self-benefitting claims (Feld & Frey, 2018). Thus, it is critical for managers, firms, regulators and policymakers to empirically know whether, when and to what extent honesty declarations can indeed curb unethical behavior (e.g., Feldman, 2017).

Among the different honesty interventions, pledges, oaths, or honor codes are the most frequently studied interventions and nudges (Hertwig & Mazar, 2022). While a recent experimental paper (Kristal et al., 2020) refuted previous claims and showed that merely signing a statement on a form (at the beginning or the end) does *not* reduce dishonesty in laboratory cheating tasks, several other laboratory studies have shown honesty pledges to be effective. These studies have focused on ex-ante pledges versus no honesty declarations and have shown that introducing an ex-ante honesty pledge can significantly reduce subsequent unethical behavior in various online and lab tasks (e.g., Beck et al., 2018; Jacquemet et al., 2019; Heinicke, Rosenkranz, & Weitzel, 2019; Peer & Feldman, 2021). For example, when participants were asked to self-report their performance in an online version of the matrix task (Mazar et al., 2008), which determined their bonus for completing the study, their reports were considerably lower (up to 50% less) if they were first asked to pledge that they will report a problem as solved only if they indeed found the solution to it (Peer

& Feldman, 2021). Other studies have also shown that honesty pledges can encourage subsequent honest responses in preference elicitation surveys (Carlsson et al., 2013; Jacquemet et al., 2013).

Particularly noteworthy are some field experiment findings that have shown effectiveness of pledges, though it is not always clear if the tested pledges were introduced to support existing ex-post honesty statements or were introduced in a setting without any honesty declarations. For example, the Social and Behavioral Science Team in the US (2015) reported on an experiment to promote more accurate self-reports of sales, and consequently the more accurate collection of fees from vendors of goods and services to the federal government. Adding an honesty pledge ("*I promise that the information I am providing is true and accurate*") at the top of the online data-entry form resulted in vendors self-reporting significantly more in sales (median amount was $445 higher) than without that honesty pledge, contributing to an additional $1.59 million in collected fees by the federal government within a single quarter. More recently, the HM Revenue & Customs (HMRC, 2019), the UK's tax administration, reported on a randomized control trial (RCT) testing the effectiveness of adding an honesty pledge before being able to fill a digital tax return form ("*I declare that the information I will give on this tax return and any supplementary pages is correct and complete to the best of my knowledge and belief. I understand that I may have to pay financial penalties and face prosecution if I give false information*."). The RCT was done with businesses filing their Value Added Tax Returns using HMRC software. According to HMRC, the 12-months long trial resulted in additional £200 million in tax revenue from 1.5 million customers for a one-off cost of £8,000.

At the same time, some studies have not been successful in reducing cheating with honesty pledges. One study found that signing an honesty declaration can even backfire: cheating rates among students taking an exam were estimated as higher for those asked to sign a pledge (Cagala et al., 2021). The authors suggest that this unexpected effect was possibly because students who signed a declaration may have believed that cheating is more prevalent than students that were not asked to sign them. Indeed, another study found that sometimes moral reminders can increase, rather than decrease, cheating, because they provide a signal that is interpreted by receivers as suggesting it is possible to cheat (Zhao, Dong, & Yu, 2019). Kristal et al., (2020) also failed to find any effect of the pledge on reducing dishonesty in the lab. Notable here may be

that these studies' choice of language (i.e., content) or presentation (i.e. the setting in which the information was conveyed) of the honesty pledges may have hampered the pledges ability to strengthen the compass for ethical behavior. For example, one of the pledges in these lab studies did not ask participants to commit to being honest or accurate but rather to give valid responses (e.g., "*Please provide your signature to certify that the information you will provide is valid.*" in Kristal et al., 2020). Additionally, a field study in Guatemala also failed to find an effect of a pledge on tax reporting (Kettle et al., 2020). In this study, taxpayers were confronted with an honesty pledge in a "CAPTCHA" pop-up window that they needed to attend to, in order to be able to proceed to filling out their online tax return forms. The authors speculated that the null effect of the honesty pledge may have been due to taxpayers perceiving the pledge as too disconnected from the actual tax return form. They also speculate that for some taxpayers the honesty pledge may have been perceived as part of the actual CAPTCHA and thus, did not receive attention in a bid to progress to the main form. Further supporting this latter hypothesis is the fact that not only did the honesty pledge not have an effect, but none of the authors' other behavioral interventions that have been shown to be successful elsewhere in improving truthful self-reports (e.g., public good message) and were added in separate treatments to the CAPTCHA pop-up window had any effects.

Together these findings suggest that while introducing ex-ante honesty pledges can be a useful tool to support ethical conduct in contexts without any other honesty declarations, it is unknown how sensitive their effectiveness is to different operationalizations. From a practical and policy perspective it is important to understand what aspects moderate an honesty pledge's ability to reduce self-benefiting misreporting. For this, it is critical to first distinguish so-called honesty statements versus pledges. While honesty *statements* are typically requested ex-post (e.g., after one has calculated their taxes or filled out an insurance claim form) and are used to remind individuals, before they submit their form, that their report must be accurate, truthful, and complete, honesty *pledges* are used ex-ante, before the relevant behavior is expected, and are designed to enhance individuals' commitment to behave ethically going forward. In our research, we focus only on the potential effects of such ex-ante honesty pledges.

Ex-ante honesty pledges could prevent or reduce unethical behavior through several potential mechanisms. First, honesty pledges may simply remind people that their report might be inspected ex-post, or that there might be negative consequences (e.g., penalties or fines) if caught having been dishonest. Another, more psychologically oriented mechanism is that once one indicates to agree with a pledge, that pre-commitment creates a moral obligation to keep one's promise to behave ethically (e.g., Wilkinson-Rayn & Baron, 2009). In addition to moral appeal, the pre-commitment to behave ethically might also influence behavior because pledges appeal to people's inherent desire to be coherent and act in self-consistent manners (Baca-Motes, Brown, Gneezy, Keenan & Nelson, 2012; Swann, Rentfrow, & Guinn, 2003; Swann & Buhrmester, 2012). Yet another mechanism through which honesty pledges have been proposed to operate is that honesty pledges may reduce individuals' ability to morally disengage when subsequently facing an ethical dilemma, thereby making it "harder" for individuals to report or act dishonestly (Bandura 1989, 1990). That is, honesty pledges may make ethics and ethical conduct more salient (i.e., harder to dismiss) or disambiguate for individuals what is it that they are expected to do and what specific behavior would be seen as the ethical thing to do, such that unethical behavior becomes harder to dismiss and/or justify (e.g., Boussalis, Feldman, & Smith, 2018; Dana, Weber & Kuang, 2007; de Bruin, 2016; Mulder, Jordan & Rink, 2015; Pittarello et al., 2015; Shalvi, Gino, Barkan, & Ayal, 2015).

The degree by which ex-ante honesty pledges can make ethical behavior more salient could depend on how individuals are asked to make the pledge in practice. Real life settings provide a myriad of ways of consenting to pledges, from having people only mark a checkbox that they "agree with the above" to approaches that incorporate forms of identification such as asking individuals to enter their ID or SSN number, to type or sign their names, etc. Thus, it appears that 'not all pledges are created equal' and it is unclear if effective pledges need to be carefully designed and if so, how. Because the research studies thus far employed different operationalizations of pledges and tested them in different settings and on different tasks and outcomes, it is impossible to compare the results across the different studies and draw more general, practical conclusions. This shortcoming calls for a more systematic examination of the effects of varying operationalizations of ex-ante honesty pledges on dishonesty. The aim of this research is to advance

this goal. We do so by exploring the contribution of two specific factors to the effectiveness of pledges: (1) the level of *identification* required from individuals to commit to the pledge and (2) the degree of effort or *involvement* when making the pledge.

Identification and involvement in honesty pledges

Legally, pledges should be "signed" by the intended signee not only to express consent to the presented terms but also to be binding and enforceable. Thus, a signature represents a mark of approval and commitment. Contrary to common belief, a legally binding signature can be as simple as marking a box "I agree with the above." That is, to be enforceable, pledges do not require to handwrite one's name, initials, or some other personal signature.

Enforceability should, rationally, deter people from providing false reports in fear of increased enforcements and sanctions (e.g., Teodorescu et al., 2021) and reduce or prevent dishonesty and unethical behaviors. At the same time, a large body of research on behavioral ethics has shown that people often decide whether to act dishonestly not only based on a cost-benefit analysis of the potential external or "real" consequences but also based on more internal or psychological consequences (e.g., Mazar et al., 2008). In particular, from a psychological perspective, for a pledge to be effective it should make salient both ethics as well as one's self such that, when faced with a temptation to misbehave, individuals will experience a strong ethical dissonance (Barkan, Ayal, & Ariely, 2015) that will curb their unethical behavior. And indeed, research on "external (non-)anonymity" and ethical behavior suggests that even if the experimenter has no way to know if a particular individual was dishonest, individuals are only more likely to engage in self-benefitting dishonesty (i.e., antisocial behavior and rule breaking) when their identity, for example, their name and address, is unknown to the experimenter (Nogami & Takai, 2008). Other research on "internal (non-) anonymity" suggests that handwriting one's name versus typing one's name is a stronger self-identity prime, leading to greater engagement with self-relevant behavior (Kettle & Häubl, 2011).

Thus, we posit that one of the necessary requirements for an honesty pledge to be effective is that it must be engaging enough to capture people's attention and change their mindset toward ethical conduct; and while asking for a commitment to a pledge through a signature requires some engagement, depending

on its operationalization (i.e., how identifiable it is), that by itself may not result in sufficiently high levels of engagement with the actual content of the pledge to make ethical conduct salient. That is, while signing a pledge may satisfy the legal requirement to allow the enforcement of sanctions in case of violations ex-post, it may not be enough for individuals to feel their self is implicated, and to ultimately prevent them from acting dishonestly going forward.

The current research

The empirical evidence about the effectiveness of honesty pledges has been mostly positive but not systematic enough to inform the implementation of pledges as means to prevent dishonesty. We posit that this is a result of a lack of a framework of how honesty pledges could be operationalized, which has led researchers and practitioners to implement pledges ad hoc in a variety of forms. We contribute a more systematic examination (with replications) of some of the elements that can make a pledge effective or not. Specifically, we examine if and to what extent the level of involvement with the content of a pledge and the requirement for identification can reduce unethical behavior of over-reporting one's performance to increase financial gains. Our general hypothesis is that enhanced involvement and identification will increase the effectiveness of pledges to reduce participants' over-reporting (i.e., cheating). Furthermore, we explore whether increasing involvement with the pledge is more effective than merely asking for identification and whether involvement and identification have different effects in the short vs. longer term.

We conducted a series of four online studies to examine these questions. In Study 1 we find that honesty pledges that require a form of involvement (i.e., manually copying the text of the pledge) outperform standard honesty pledges that require only reading the pledge (which does not reduce dishonest reporting). In Study 2 we find that a request for a more personal, albeit still anonymous[1] signature (operationalized as handwriting or typing one's first name or initials), increases the effectiveness of the pledge to be comparable to copying it. In Study 3 we replicate these results on a larger sample and also

[1] Due to ethical concerns, we do not explore in this research pledges that require full names or other personally identifying information.

show that attention to the content of the pledge is increased when asked to copy it. Lastly, in Study 4 we examine the effect of honesty pledges over time on two consecutive tasks, with a short delay between them, and find that the effect of copying the pledge lasts longer over time, while that of only reading the pledge dissipates. Finally, we also find that repeating the pledge is important to counteract the observed decay of its effectiveness over time. We report all data exclusions, all manipulations, and all measures in all the studies. All the studies were pre-registered using AsPredicted.org and these are available, along with all data and research materials at https://osf.io/hda7b/?view_only=4ce2fa3bf29c4df1adb532aa1faf3dd6 .

Study 1 – Identification and Involvement

The first study examined the effects of four different honesty pledges that included a request for identification, or not, and asked for either low or high effort (i.e., involvement) in making the pledge, and compared these to a condition without any pledge.

Method

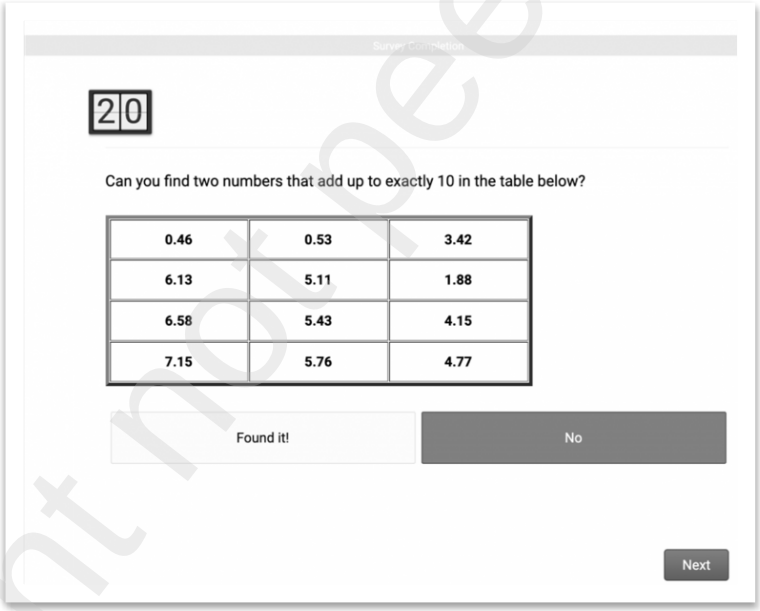
Participants. Nine hundred and one participants, recruited from the United States on Prolific, completed the study. We excluded 25 participants who failed to follow instructions (based on our pre-registered criteria). The final sample (N=876) included 47.9% females (50.7% males, 10 identified as "other" and 2 declined to disclose), with a mean age of 32.24, (SD=10.85). Participants were paid 0.5 GBP for completing the study and an additional bonus according to their reporting, as detailed in the following.

Design and procedure. Participants were invited to a study about problem solving with an opportunity to gain an additional bonus of up to 2 GBP. First, similar to the design in Peer & Feldman (2021), participants were given instructions about the task. They were told that they would be asked to solve multiple short problems involving simple calculations. Specifically, their task would be to find the two out of twelve numbers in a matrix that, when added up, result in exactly 10 (as in Mazar et al., 2008) in 20 seconds or less (see Figure 1). Participants were asked to summarize, briefly (with at least 50 characters) and in their own words, the instructions of the task. Then, participants were asked to complete a practice trial (which was identical for all participants) and see if they could find the solution to the problem in 20 seconds or

less. Finally, participants were given additional instructions according to the condition to which they were randomly assigned.

In the Control condition, participants were told that they would go through 20 problems and earn 0.1 GBP for each problem they solved correctly. That is, they learned that for each problem for which they indicated they found the solution (i.e., marked "Found it!"), they would be asked, on a subsequent page, to enter the two numbers that they had found to add up to exactly 10. If they provided the correct solution, they would earn 0.1 GBP per problem. If they were incorrect, they would not get the bonus for the problem. Participants were also told that there were no penalties for incorrect answers. Then, participants were asked to start the matrix task when they were ready.

Figure 1. Example of a problem in the Online Matrix Task.



Note: Participants had 20 seconds of time to actively click the "Found it!" button to gain a bonus for the problem. After 20 seconds, if no action had been taken, a "No" response was recorded and the survey page automatically advanced to the next problem.

In the other experimental conditions, participants were told that they would go through 20 problems and earn 0.1 GBP for each problem for which they indicated they found the solution (i.e. marked "Found it!"). In the Self-Report condition, participants then proceeded to the matrix task. In the other four conditions, participants were first asked to commit to a pledge (text adopted from Peer & Feldman, 2021)

before starting the matrix task, in one of the following manners (between-subjects). In the standard "Read" condition, participants were asked to read the text of the pledge ("*I promise that I will only report to have a solution to a matrix problem after verifying carefully that indeed I have found two numbers that add up to 10. I know that I will be paid based on my reporting and hence will take it very seriously to be accurate in my reporting.*") and then mark a check box saying "I agree" to declare they agreed with the text of the pledge. In the "Read + ID" condition, participants were asked to read the same text of the pledge, and then enter their Prolific Participant ID to declare they agreed with the pledge. In the "Copy" condition, participants were asked to manually re-type the text of the pledge into an open text box to declare they agreed with it (the text of the pledge was presented as an image file to prevent copy-and-paste). In the last, "Copy + ID" condition, participants were asked to re-type the text of the pledge and to enter their Prolific Participant ID. Thus, these four conditions manipulated the levels of identification (low: Read or Copy vs. high: Read + ID or Copy + ID) and involvement (low: Read or Read + ID vs. high: Copy or Copy + ID). Together, with the Control (where participants had to provide a solution for each problem and were only paid for correct responses) and the Self-Report condition (where participants did not have to provide the actual solutions and therefore over-reporting lead to a higher bonus), the design included six between-subjects conditions.

The order of the 20 problems was set randomly in advance and was identical for all participants across all conditions. Out of the 20 problems, three (#5, #10, #16) were unsolvable as they did not include any combination of two numbers that added up exactly to 10. After completing the 20 problems of the task in their respected condition, participants were asked to report their age, gender, income (household income in USD before taxes last year) and level of education. Participants were also asked whether they recalled doing a study like this one in the past and could add any comments they had before submitting the study to receive payment. Participants in the Control condition were paid according to the number of correct solutions they provided (as checked by an RA) and participants in the other conditions were paid according to the number of problems they self-reported as solved.

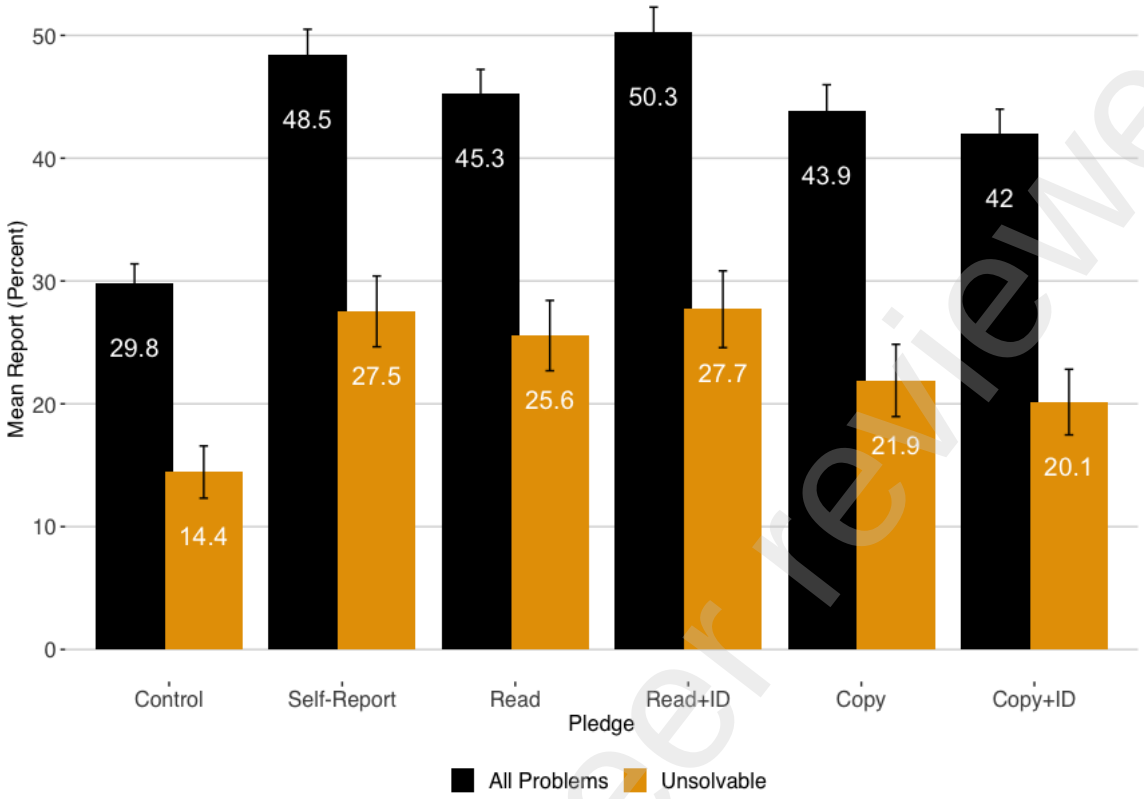
Results

The main dependent variable was the percent of the total number of available problems (i.e., 20 matrixes) that participants actually solved (in the Control condition) or self-reported (in all other conditions) as solved[2]. Using G*Power software (version 3.1), a sensitivity analysis with alpha = 0.05 and power = 0.80, we found that this sample size is sufficient to detect a main effect size of $f = 0.12$ between the five conditions. ANOVA showed that the report levels varied significantly between the six conditions, F (5, 870) = 13.94, $p < .001$, $\eta^2 = 0.08$. As can be seen in Figure 1, participants' report was lowest in the Control condition (29.77% of problems actually solved), and significantly higher in the Self-Report condition, when participants had the opportunity to cheat for higher pay (48.45%, $M_{diff} = 18.68$, t (290.66) = 7.15, $p < .001$, 95% CI_{diff} [13.54, 23.83], see Supplementary Table S1 for the mean, N, and SD for all conditions)[3].

Figure 2. Mean percent of problems reported as solved, across all problems or across unsolvable problems only, between conditions in Study 1.

[2] Note, that means that in studies with a subset of unsolvable matrixes, Control condition participants could not reach 100%; while participants in all other conditions could reach 100%, if being maximally dishonest.

[3] Comparing the mean of Self-Report condition to the percent of problems *claimed* as solved in the Control condition (M=37.20, SD=21.26) showed a similar effect, t (296.82) = 4.198, $p < .001$.



* Error bars show one standard error above/below the mean.

In comparison to the Self-Report condition, three of the four pledge conditions reduced reported performance: while the mean difference between the Self-report and Control conditions was 18.68, it was reduced to 15.5 when participants were only asked to read the pledge, reduced further to 14.09 when participants were asked to copy the pledge, and reduced the most to 12.21 when participants were asked to both copy the pledge and enter their ID. These differences show a relative reduction in over-report (compared to the baseline over-report, which is the mean difference between the Self-Report and Control conditions) of 17%, 25%, and 35%, in the Read, Copy, and Copy + ID conditions, respectively. The Read + ID condition actually led to an increase of 10% in mean reports. Statistically analyzing the differences in report levels between each of the four pledge conditions to the Self-Report condition (without the Control condition) showed that only the Copy + ID conditions reduced over-reports significantly, t (291.52) = 2.26, p = .012; all other pairwise differences not significant, see details in Table S1).

Unsolvable problems. We next examined the percent of unsolvable problems that participants claimed to have found a solution for. As seen in Figure 1 (and detailed in Table S1), participants in the

Self-Report condition claimed more unsolvable problems as solved than participants in the Control condition, t (281.15) = 3.65, $p < .001$. Compared to the Self-report condition, only the Copy + ID condition showed a significant reduction in the percent of unsolvable problems reported as solved, t (291.88) = 1.88, $p = 0.031$.

Discussion

The results of Study 1 show that pledges can, under certain conditions, reduce people's propensity to over-report in order to increase their gains (i.e., cheat), which is consistent with previous studies (e.g., Beck et al., 2018; Jacquemet et al., 2019; Peer & Feldman, 2021). Among the different forms of pledges, it appears that ensuring participants' involvement by asking them to exert effort and copy the text of the pledge (as in Peer & Feldman, 2021) and to add their ID, had the strongest effect in increasing the chances that participants would comply with the pledge and curb their dishonest behavior. Notably, asking to copy the pledge without providing an ID (Copy) resulted in lower absolute levels of cheating than asking to read the pledge and provide an ID (Read + ID); suggesting that involvement (copying the pledge) played a larger role than identification in the effect of the significant Copy + ID condition. However, one can claim that providing a participant ID is not a very strong means of identification, as it still anonymous and may not hold much personal relation such as providing a name, signature or initials may hold. Moreover, previous studies showed that if awareness to the self is increased (e.g., by a message of "don't be a cheater") it can reduce unethical behavior (Bryan, Adams, & Monin, 2013). Thus, in the next study, we aimed to re-examine these findings using such other forms of identification to examine whether more enhanced means of identification may also be effective in reducing dishonesty.

Study 2 – Enhanced Identification

Method

Participants. We recruited 1,401 participants from the United States on Prolific, and excluded 27 that did not follow instructions (as pre-registered). The final sample (N=1374) included 46.6% females (51.9% males, 17 identified as "other" and 4 declined to disclose), with a mean age of 32.33 (SD=11.0). Participants were paid 0.5 GBP plus a bonus of up to 2 GBP.

Design and procedure. The procedure was identical to Study 1 except that we had participants solve 10[4] problems and they could receive a bonus of 0.2 GBP per problem. As in Study 1, participants were randomly assigned to either a Control condition (with no option to cheat for higher bonus), Self-Report, or to variations of pledge conditions. Four of the five pledge conditions asked participants to read the pledge and to provide an identification in four different forms: two of the forms asked for a signature (sign with your first name or sign with your initials), one asked to type in your initials, and the last to type in your Prolific Participant ID. The fifth pledge condition asked participants to copy the text of the pledge (without any identification) like in Study 1. The order of the problems was pre-randomized and fixed between conditions, and two problems (matrix #2 and #8) was unsolvable.

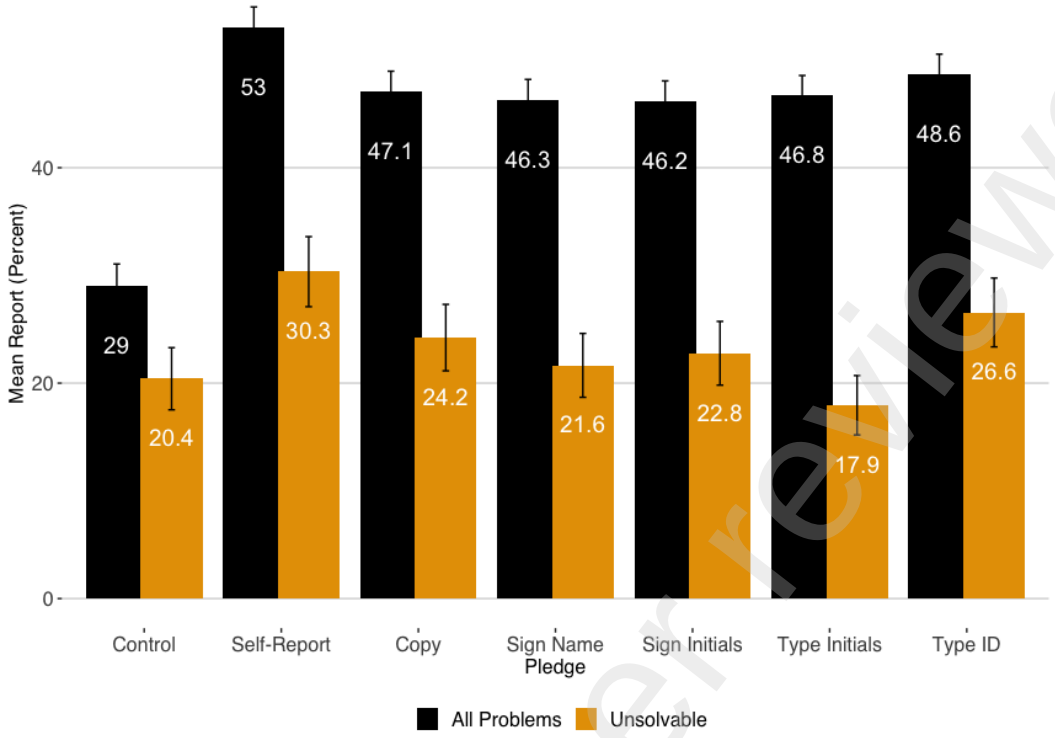
Results and discussion

Sensitivity analysis with alpha = 0.05 and power = 0.80 showed that sample size is sufficient to detect a main effect size of $f = 0.10$ between the seven conditions. As seen in Figure 3 (and detailed in the Supplementary Table S2), we found statistically significant differences between the seven conditions in the percent of problems reported as solved, F (6, 1320) = 13.42, $p < .01$. As expected, participants' reports were lowest in the Control condition (29% of problems solved), and highest in the Self-Report condition, when participants could cheat (53.03%, t (328.71) = 8.59, $p < .001$), a mean difference of 24.04, which is regarded as the baseline over-report (cheating) degree[5].

Figure 3. Mean percent of problems reported as solved, across all problems or across unsolvable problems only, between conditions in Study 2.

[4] In the pre-registration of this study, there was a mistake as in one instance it said that the task included 20 problems (but earlier in the document it correctly states it had only 10 problems). As the materials on OSF show, there were indeed only 10 problems in the task in this study.

[5] The difference between Self-Report and the percent of problems *claimed* as solved in the Control (M=42.40, SD=27.08) was also significant, t (394.46) = 3.94, $p < .001$.



* Error bars show one standard error above/below the mean.

Asking participants to copy the text reduced over-reports to a mean difference of 18.12 to the control (-25% reduction compared to the difference between Self-Report and Control). In contrast, asking participants to only type their ID reduced over-reports less to a 19.66 mean difference (-18% reduction). Asking participants to sign their name or sign their initials reduced over-reports to 17.3 and 17.2, respectively (-28% reductions), and asking participants to type initials showed a similar effect (-26% reduction). The differences between these pledge conditions and the Self-Report condition were statistically significant (see Table S2), except for the Type ID condition that did not significantly reduce cheating.

Unsolvable problems. As seen in Figure 3 (and detailed in Table S2), all pledges led to a reduction in the percent of (the single) unsolvable problem reported as solved, compared to the Self-report condition, in a similar pattern as above. However, these differences were not statistically significant, $p > .05$.

To summarize, similar to the results of Study 1, we found that asking to copy the pledge (Copy), reduced cheating significantly, and asking participants to only enter their Prolific ID (Type ID) was ineffective at reducing cheating, as in Study 1 (and as in Peer & Feldman, 2021). However, we found that

other, more enhanced forms of identification that involve signing or typing names or initials reduce cheating to a degree similar to (i.e., not less than) that of the Copy condition. Thus, it appears that while involvement seems to produce the desired effect of the pledge more consistently, identification may do so also, but only when it includes a personal piece of information (initials or name – typed or signed) and not when it only includes an anonymous ID number. This suggests that common forms, which ask individuals only to check a box or mark their agreement with a veracity or honesty statements might not be effective in reducing cheating, as indeed was found in the field study mentioned earlier (Kettle et al., 2020). In contrast to previous lab studies (Kristal et al., 2020) that did not find an effect for a signature, we did find that asking for a signature ex-ante reduced cheating significantly and considerably, suggesting that enhancing the identification required in a pledge is important.

In the next study, we aimed to validate and replicate these findings. Additionally, we test the combination of a high involvement pledge with a personal signature element in order to understand whether they operate through a similar mechanism or would their effects be additive and, when combined, reduce cheating the most. To further explore the mechanisms underlying the different effects of different pledges, we also included two additional measures in the next study. First, we explored whether effects could be the result of increased attention to the content of the pledge, predicting that high involvement will result in higher attention to the content of the pledge, which will manifest in better recall of the content of the pledge after it was made. This increased attention, we conjectured, might explain the larger effect of the high involvement pledge. As another potential mediator, we also explored how the different pledges elicit concerns of being sanctioned for cheating.

Study 3 – Signature and Copy Effects

Method

Participants. We recruited 1,498 participants from Prolific (51.2% female, 47.5% males, 12 identified as "other" and 7 declined to disclose their gender; Mean age = 36.9, SD=11.9). Following our pre-registered procedure, this sample did not include participants who said they completed a similar study in the past

(N=260) or failed an attention-check question at the beginning of the study (N=319), or did not follow instructions of the pledge (N=3).

Design and procedure. After reading the instructions and completing a practice task, participants were asked to solve 10 matrix problems. We reduced the time limit for each matrix to 10 seconds, and half of the problems (the even numbered ones) were unsolvable. This study had five conditions that included a Self-Report condition and four pledge conditions: Copy, Read, Copy + Sign name, Read + Sign name. We did not include a control condition in this study as our main focus was on how the different pledges compare to the Self-Report condition, when no pledge is included.

We also measured attention to the pledge by asking participants, at the end of the study, to identify, out of six quotes (three correct and three incorrect), which ones appeared in the text of the pledge (similar to the procedure in Shu et al., 2011). The score was computed by giving +1 for every correct quote that was marked, and a -1 for every incorrect quote that was marked, yielding a score ranging from (-3) to (+3). Additionally, we also explored whether the effect of the pledges could stem from differences in the perceived fear of being sanctioned by the researchers for cheating. This included three items to which participants were asked to indicate their agreement from 1 (strongly disagree) to 5 (strongly agree): "I thought that reporting a problem as solved, when I actually did not solve it, might lead to my submission being rejected by the researcher", "I felt that if I report solving too many problems, the researcher will not believe me and might reject my submission", "I was sure that my submission will be approved, regardless of how many problems I report as solved" (reverse coded).

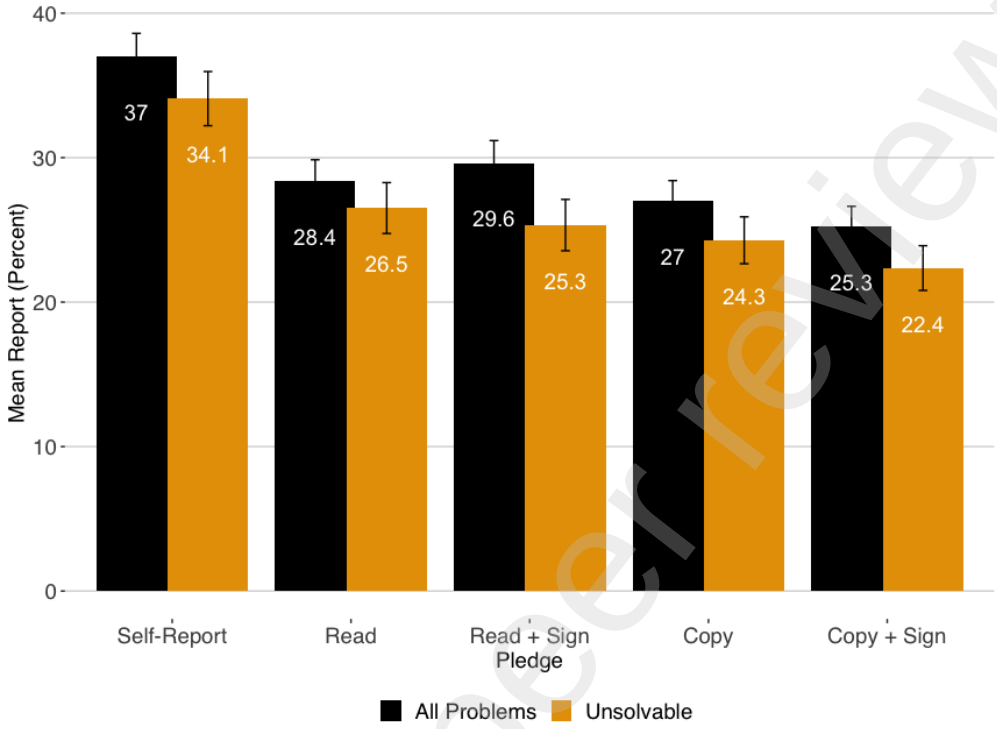
Results and discussion

Sensitivity analysis with alpha = 0.05 and power = 0.80 showed that sample size is sufficient to detect a main effect size of $f = 0.09$ between the five conditions. We found significant differences between conditions in the percent of problems reported as solved, both for all problems, F (4, 1493) = 9.34, $p < .01$, or for the unsolvable problems, F (4, 1493) = 6.84, $p < .01$. As can be seen in Figure 4, all four types of pledges significantly reduced participants' reported performance in comparison to the Self-Report condition (see full details in Supplementary Table S3). Compared to the Self-Report condition, asking

participants to only read the pledge reduced reports by 23% while asking to read and sign by name reduced by 20%. Asking participants to copy the pledge reduced reports by 27% and asking to copy and sign by name reduced reports in the highest degree of 32%, compared to the Self-Report condition. All of these pairwise difference were significant ($p < .001$, see Table S3).

Analyzing the differences between the four pledge conditions, using a 2 X 2 (Involvement: Read or Copy vs. Identification: Sign Name or not) ANOVA showed a significant effect only for involvement, F (1, 1193) = 3.88, p = .04, but not for identification nor their interaction, F (1, 1193) = 0.03, 1.09, respectively and p = 0.86, 0.29, respectively. The averaged standardized difference (Cohen's d) between the two Copy conditions to the two Read conditions was small, d = 0.11 (95% CI [0.01, 0.23]).

Figure 4. Mean percent of problems reported as solved, across all problems or across unsolvable problems only, between conditions in Study 3.



* Error bars show one standard error above/below the mean.

Explored mediators. We examined attention to the pledge's content by asking participants to recall which direct quotes appeared in the text of the pledge (a measure that ranged from (-3) to +3, as explained above). We found that the percent of participants with a recall score above zero (indicating some accurate recollection of the pledge) was 69.4% and 64% in the Read and Read + Sign Name conditions, respectively, while it was 95.3% and 93.6% in the Copy and Copy + Sign Name conditions, respectively, x^2 (3) = 150.01, $p < .001$. Similar differences were found using a higher threshold of a recall score above 1 (29.9%, 19% for Read and Read + Sign Name, respectively vs. 77.6%, 68.7% for Copy and Copy + Sign Name, respectively, x^2 (3) = 295.9, $p < .001$). ANOVA on the continuous recall score showed significant effects for both Involvement and Identification, F (1, 1193) = 477.36, 13.94, respectively, $p < .001$, but no significant interaction, F (1, 1193) = 0.08, $p = 0.771$. The effect size of Involvement on recall was large, Cohen's d = 1.26 (95% CI [1.13, 1.38]), while the effect of Identification was negative and small, d = -0.18 (95% CI [-

0.07, -0.29]). However, a mediation analysis did not show a statistically significant indirect effect of Involvement on reports through the level of recall, 95% CI [-0.31, 0.05].

Regarding fear of sanctions, Cronbach's alpha for these items was too low (0.51) in order to average the three items to one score. We did not find any significant differences between the pledges in how participants responded to any of the three items, F (3, 1193) = 0.2, 0.95, 0.91 p = 0.897, 0.416, 0.434, respectively.

The results of this study suggest that all these types of pledges can be effective in reducing dishonest reporting, and that increasing involvement (by a request to copy the pledge) adds a significant, albeit small marginal increase to the effect of the pledge. Relatedly, people are more attentive to the content of the pledge when they are asked to copy it, but the study did not find empirical support for a mediation of the level of attention (measured by recall) on the effectiveness of the high-involvement pledges on dishonesty.

Study 4 – Effects over Time

The three studies thus far examined the short-term effect of pledges. That is, how does a pledge made by participants right before the opportunity to cheat (in the matrix task) affects their behaviors. In many real-life situations, however, people are routinely asked to make pledges before being asked to make several sequential or separate reports, each of which may provide a temptation to cheat. Previous research has not yet examined how broad (or narrow) the effect of a pledge can be over time and whether its effect will decay or persist when different opportunities to cheat arise after making the pledge. Relatedly, it would be important to understand how reminding the pledger of their pledge may strengthen the effect of the pledge and perhaps prevent any decay over time. A recent study examined this question tangentially with participants being given reminders of their pledge in the middle of a task (Le Maux & Necker, 2023). However, that study involved several conditions in which the pledge was preceded with different types of moral reminders. In some conditions, the researchers found that repeating the pledge increased honesty afterwards but in other conditions it did not change or even reduced honest reporting. Thus, it is unclear how and when reminding or repeating a pledge would be important.

Additionally, in many instances, there could be some amount of time that passes between when the pledge is made and when the temptation to act unethically actually arises. The question of whether pledges can still exert an effect when the opportunity to cheat is distant from the time the pledge was made is unanswered even though it is important both from a theoretical perspective (to understand the scope and boundaries of pledges' effects) and from a practical perspective (to inform policy makers how to design the pledge process or when to expect meaningful effects of pledges). Pledges, like many other instructional tools, could suffer from habituation or adaptation effects if they are used too often or for too many purposes (e.g., Ben-Shahar & Schneider, 2014). This habituation and adaptation to (un)ethicality (i.e., when people promise to behave ethically and fail to honor their promise), which can be anticipated psychologically across many different mechanisms (Bandura, Barbaranelli, Caprara, G, & Pastorelli, 1996), presents a major challenge to the use and effectiveness of pledges in real-life settings.

Thus, in our last study we aimed to explore the effects of different types of pledges over longer periods of time, and also to examine whether the effects of pledges we discovered in the previous studies may persist (or decay) if there is some delay between the time individuals make the pledge to when they are faced with the opportunity to cheat (in the task). Lastly, we also examined whether asking people to repeat the pledge would mitigate any decay in their effectiveness. To examine these questions, participants in Study 4 completed two rounds of the matrix task, with a short delay between them, and some of the participants repeated the pledge after the delay or not. Because the previous studies showed a consistent difference between pledges of high vs. low involvement (i.e., Copy vs. Read), we focused Study 4 on these two types of pledges.

Method

Participants We recruited 1,505 participants from Prolific (49.4% females, 49.5% males, 12 identified as "other" and 4 declined to disclose; mean age was 38.4, SD=13.2). Following our pre-registration, we precluded participants that completed a similar study (on our Prolific account) but did not exclude any participants from the final sample. Participants were paid 1 GBP plus a bonus of up to 3 GBP.

Design and procedure. After receiving instructions and completing a practice task, as in the previous studies, participants were asked to solve 15 consecutive matrices for a bonus of 0.1 GBP each. Afterwards, participants were asked to view a short neutral video (a TEDEd talk by Elizabeth Cox on "The Benefits of Daydreaming", about five minutes long, participants could not proceed before the video time elapsed) and answered two attention check questions about it[6]. Then, participants were asked to solve another set of 15 matrices for an additional bonus of 0.1 GBP each.

Participants were randomly assigned to one of five conditions: one Self-Report and four pledge conditions (Copy, Copy + Copy, Read, Read + Read). In the Self-Report condition, participants did not encounter any pledge and were paid based on the number of problems they claimed having solved (allowing to cheat for higher pay). In the two Copy conditions, participants were asked either once (Copy: before the onset of the first set of 15 matrixes) or twice (Copy + Copy: before the onset of each of the two sets of 15 matrixes) to consent to a similar pledge[7] as used in the previous studies by retyping it. In the two Read conditions participants were asked either once (Read: before the onset of the first set of 15 matrixes) or twice (Read + Read: before the onset of each of the two sets of 15 matrixes) to mark a checkbox that they agreed with the presented pledge.

Results and discussion

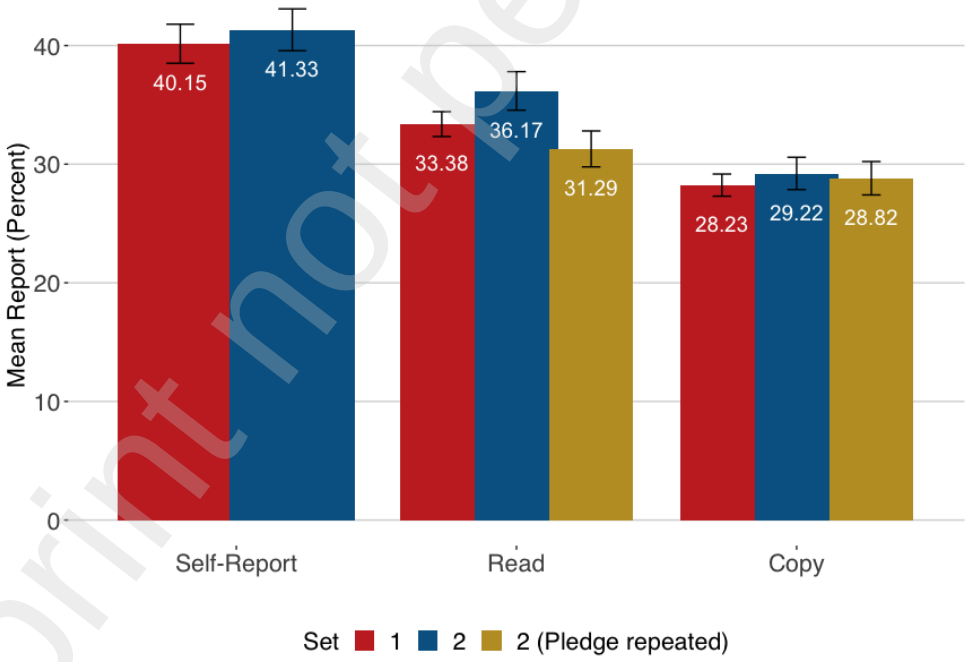
Sensitivity analysis with alpha = 0.05 and power = 0.80 showed that sample size is sufficient to detect a main effect size of $f = 0.09$ between the five conditions. In the first set of problems, participants in the Self-report condition reported an average of 40.15% (SD=28.8) of the problems as solved. In the Read condition, participants reported less problems solved (M=33.38%, SD=25.9) and in the Copy condition the reports were even lower (M=28.23%, SD=22.7), F (3, 1502) = 22.65, $p < .001$. Because the two repeat

[6] About 14% answered the video attention check questions wrong. However, we did not find any significant interaction between these questions and the pledge conditions either on the second trial or overall, $p > .06$, and thus we do not consider this variable in our analyses.

[7] "I promise that I will only report a solution to a problem after verifying carefully that I have indeed found two numbers that add up to 10. I know that I will be paid based on my reporting and hence I will take it very seriously to be accurate in my reporting."

pledge conditions (Copy & Copy, Read & Read) were identical to the two non-repeat pledge conditions (Copy, Read) for the first half of the study (i.e., through Set 1 and the video), we collapsed across the repeat and non-repeat conditions for Set 1, but not for Set 2. In the second set of problems, participants in the Self-Report condition reported solving 41.33% on average (SD=30.8). In the Read condition, without repeating the pledge, participants reported 36.17% (SD=28.4) of problems, and 31.29% (SD=26.36), with repeating the pledge. In the Copy condition, without repeating the pledge, participants reported 29.22% (SD=23.2) of problems, and 28.82% (SD=24.42), with repeating the pledge, F (4, 1500) = 11.93, $p < .001$. Figure 5 shows these differences in the first set (between three conditions according to the pledge they were asked to make before the first set) and in the second set (between five conditions including whether the pledge was repeated or not).

Figure 5. Mean percent of problems reported as solved between conditions and sets in Study 4.

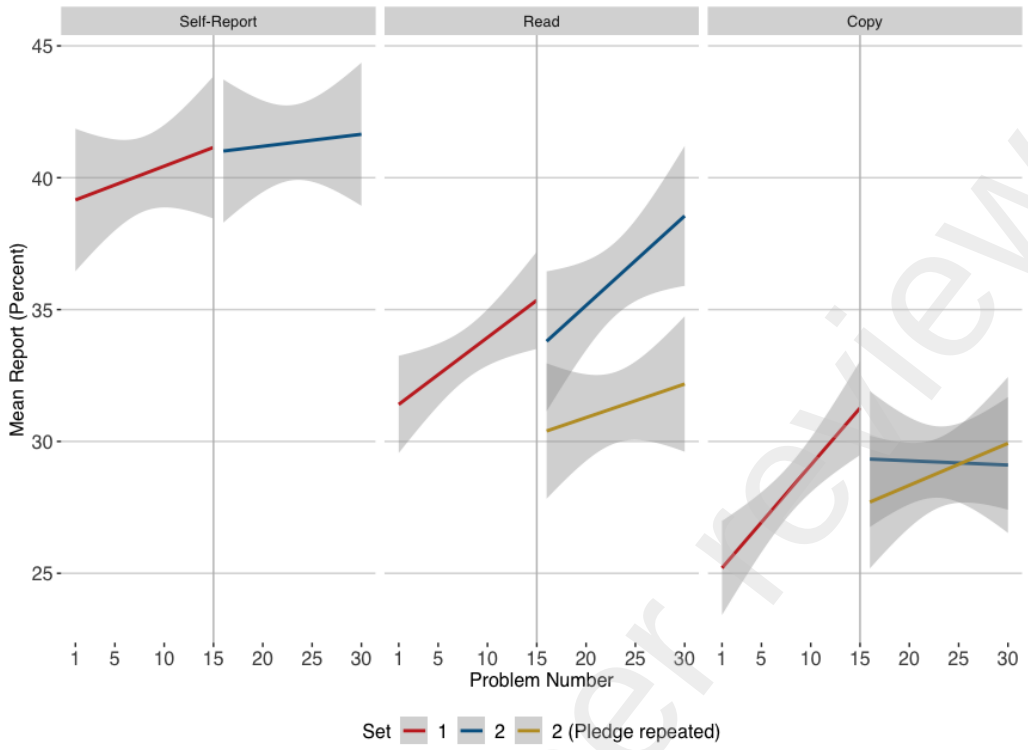


* Error bars show one standard error above/below the mean.

To examine the effects of the different pledges over time, we conducted another analysis, that was not pre-registered, on the mean reports over the sequential problems and across sets. Figure 6 shows that

when no pledge was administered (Self-Report condition), the mean report of problems solved slightly increased over time and between the first and second matrix set, but this trend was not statistically significant, $b = 0.14$, SE = 0.17, $p = 0.396$. In the Copy condition, the mean report was considerably smaller and showed a significant increase over problems, $b = 0.49$, SE = 0.16, $p = .002$. However, there was also a significant interaction of problem number with matrix set, $b = -0.51$, SE = 0.22, $p = 0.022$, indicating that this trend discontinued in set 2: As seen in Figure 6, reports did not increase across problems in set 2. There was also no significant interaction between problem number and repeating the pledge, $b = -0.12$, SE=0.22, $p = 0.578$, suggesting the trend over time was similar in set 2 whether the pledge was repeated or not. For the Read pledge condition, reports started higher in the first matrix set than in the Copy conditions and increased significantly through both sets, $b = 0.52$, SE=0.16, $p = 0.001$ up to a point that the mean report on the last problems in set 2 was almost similar to that of the Self-Report condition. However, there was a significant interaction between problem and repeating the pledge, $b = - 0.49$, SE = 0.23, $p = 0.033$, showing that when the Read pledge was repeated before set 2, mean reports were lower across problems.

Figure 6. Mean report of problems as solved across problems and sets, between conditions in Study 4.



Attention. As in the Study 3, we again found higher attention (recall) rates among those who had to Copy the pledge: 89.3% received a score greater than zero (indicating they identified quotes from the pledge correctly) when they made the pledge once (only before the first set), and 96% when the pledge was repeated (again before the second set). In contrast, in the Read condition only 55.4% recalled quotes from the pledge when it was made once and 67.7% when the pledge was repeated. These differences were statistically significant, x^2 (3) = 180.89, $p < .001$. ANOVA on recall levels showed significant effects for the type of the pledge and for whether it was repeated, F (1, 1195) = 486.2, 77.76, $p < .001$, but not for their interaction, F (1, 1195) = 2.261, $p = 0.133$, confirming that copying the pledge and repeating it leads to highest levels of recall and attention to the content of the pledge. Furthermore, the effect size of copying the pledge on recall rates was larger (Cohen's d = 1.23, 95% CI [1.11, 1.36]), than the effect size for repeating the pledge (Cohen's d = 0.44, 95% CI [0.32, 0.56]).

To test whether attention (recall) mediated the effect of the different pledges, and the effect of repeating the pledge on that mediation, we conducted a moderated mediation analysis (Hayes, 2015; Model 7) with type of pledge (Copy vs. Read) as the independent variable, recall level as the mediator, whether

the pledge was repeated as the moderator, and the number of problems reported as solved in Set 2 (reports) as the dependent variable. We found that the indirect effect of recall on reports was not significant either when the pledge was repeated (b = -0.08, 95% CI [-0.35, 0.17]) or when it was not repeated (b = -0.09, 95% CI [-0.38, 0.19]. The index for the moderated mediation (difference between the above indirect effects) was also not significant, 0.01, 95% CI [-0.2, 0.06]. Similar non-significant results were found when the dependent variable was the difference in reports between Set 2 and Set 1.

To summarize, the results of this study showed that the effects of pledges with higher involvement (the Copy condition) are stronger, compared to low involvement pledges (the Read condition), as was found in our previous studies. This difference was significant not only in the short term, but also for the longer span across sequential tasks and over time, extending the findings of the previous studies. Repeating the pledge after a (short) delay contributed to increasing its effectiveness, but mostly when the pledge was initially of low involvement (Read). When the pledge already included higher involvement (Copy), its effect persisted after the delay when it was repeated or not. Attention to the content of the pledge, measured by recalling its text, was considerably higher when asked to copy the pledge, but this effect did not seem to mediate or explain the higher effectiveness of the pledge with higher involvement on reducing cheating, as was also found in Study 3.

General Discussion

Honesty pledges attempt to change the level of ethicality in peoples' subsequent behavior (e.g., de Bruin, 2016). However, previous research on honesty pledges did not systematically determine what are the factors that make them effective and to what extent. Understanding that is important both theoretically and practically. From a theoretical standpoint, without understanding when and why pledges work, it is impossible to advance the study of pledges or propose coherent explanations for their effects on behavior. From a policy perspective, understanding the "why" pledges actually work can inform policy on the "how" to better implement them. After all, honesty pledges could be applied to many contexts in which managers, policymakers and regulators are interested in enhancing ethical behavior through trust-based means.

Our findings provide several important contributions to the existing body of knowledge on whether, when and why can honesty pledges reduce unethical behavior: We document a substantial and reliable effect of pledges that replicated across four pre-registered studies with large sample sizes. Our findings suggest that when honesty pledges are properly implemented, they can reduce as much as 15-30% of unethical behavior observed under the settings of the current studies (i.e., reporting self-performance for higher financial gains in an online context and without any external costs; Mazar et al., 2008; Peer & Feldman, 2021). However, at the same time, our findings also show that when pledges are poorly implemented – as by asking people to only mark a checkbox after presumably having read it – the effect of the pledge is smaller and often insignificant.

In that regard, our findings provide insight on when honesty pledges could have stronger or weaker effects on reducing unethical behavior. In particular, while signatures are legally required and expected to increase compliance with a pledge, across our four studies we find that it is more effective to make sure that people are involved with the pledge. Notable in particular is the finding of our Study 3, which directly contrasted involvement and identification. This insight is especially important in contexts where getting individuals' full identification ex-ante might be problematic due to privacy or logistical constrains. Furthermore, this recognition is important because in the literature, in most cases when pledges have been used, it has been done without giving enough attention to the mechanisms through which those pledges may have operated. Accounting for factors such as involvement and identification when evaluating the efficacy of pledges may help solve some of the observed inconsistencies in previous research.

However, our studies had some limitations that should be addressed in future research that will further explore and validate the effectiveness of honesty pledges. First, our experiment could not include (very) high incentives and we cannot generalize our findings to situations where the stakes in cheating are considerably larger, such as large-scale frauds. In that regard, our studies focus on how honesty pledges may curb dishonesty in everyday lives, in common situations that could apply to many normative individuals. Second, while we discovered that involvement in the pledge can be key to its effectiveness, we only tested one operative mode of making the pledge more engaging, while many others could exist that

would yield the same effect. For example, pledges can be made to be read aloud, or posted in public, or one can have the pledger formulate the content of the pledge themselves, to increase their involvement in it. Future research may examine that, and should also further examine the underlying mechanism of such pledges. Future research should also replicate and expand the reported findings to other situations, other tasks, more diverse populations and may also examine whether affidavits (which are pledges confirmed by lawyers) may have different effects than the pledges examined here. The scope of future research, which can be far reaching and could not have been covered in this paper, should be expanded and examined in order to further understand the breath and boundaries of the effectiveness of honesty pledges.

From a policy perspective, our findings recommend managers and policy makers find the most feasible and best ways to increase individuals' involvement in pledges, affidavits, oaths, contracts, or other ex-ante statements that are supposed to prevent dishonesty and ensure compliance. Honesty pledges that succeed particularly in engaging individuals could function as a "soft" form of a pre-commitment device (e.g., Bryan, Karlan, Nelson, 2010). There is much evidence showing that pre-commitment devices – typically locking-in individuals in a certain action at a considerable immediate cost – can increase people ability for self-control (Duckworth, Milkman, & Laibson, 2018; Rogers & Milkman, 2016; Rogers, Milkman, & Volpp. 2014). Thus, if designed properly, pre-committing pledges could have important and considerable benefits to society and public policy: Regulators may then relax some administrative requirements and simplify procedures for many activities such as importing goods, starting a new business, reporting taxes, applying for permits and licenses, claiming due benefits, etc. Over-complicated and excess regulation can result in welfare harm and hampered growth (e.g., Sunstein, 2020), and if some requirements could be substituted by asking for ex-ante pledges, these welfare costs could be avoided, and citizens' lives improved.

In addition, effective pledges could allow policymakers to reduce resources currently allocated for lengthy and costly checks and inspections, that also increase the time citizens and businesses have to wait for responses and focus on more effective post-hoc audits. What is more, pledges could serve as market equalizers, allowing better competition between small businesses, who normally cannot afford long waiting

times for permits and licenses, and larger businesses who can. Finally, pledges, being regulation instruments that are based on trust, may gradually improve and re-build trust between state regulators and the citizens and businesses they serve, as well as between managers to their employees or between leaders and their teams. This could, eventually, increase trust in public institutions, which can result in improved efficiency, incentivized growth, and improved welfare as well as inter-personal trust between individuals engaging in mutual contracts, in many social interactions.

We report all data exclusions, all manipulations, and all measures in all the studies. All the studies were pre-registered using AsPredicted.org and these are available, along with all data and research materials at https://osf.io/hda7b/?view_only=4ce2fa3bf29c4df1adb532aa1faf3dd6 .

References

Baca-Motes, K., Brown, A., Gneezy, A., Keenan, E. A., & Nelson, L. D. (2013). Commitment and behavior change: Evidence from the field. *Journal of Consumer Research*, *39*(5), 1070-1084.

Bandura, A. (1989). Self-regulation of motivation and action through internal standards and goal systems. In L. A. Pervin (Ed.), *Goal concepts in personality and social psychology* (pp. 19-85). Hillsdale, NJ: Erlbaum.

Bandura, A. (1990). Selective activation and disengagement of moral control. *Journal of Social Issues, 46(*1), 27–46.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology*, *71*(2), 364.

Beck, T. Bühren, C. Frank, B. and Khachatryan, E. (2018). Can Honesty Oaths, Peer Interaction, or Monitoring Mitigate Lying?, *Journal of Business Ethics* 1-18.

Ben-Shahar, O., & Schneider, C. E. (2014). More than you wanted to know. In *More Than You Wanted to Know*. Princeton University Press.

Boussalis, C., Feldman, Y., & Smith, H. E. (2018). Experimental analysis of the effect of standards on compliance and performance. *Regulation & Governance*, *12*(2), 277-298.

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

Bryan, G., Karlan, D., & Nelson, S. (2010). Commitment devices. *Annu. Rev. Econ.*, *2*(1), 671-698.

de Bruin, B. (2016). Pledging integrity: Oaths as forms of business ethics management. *Journal of Business Ethics*, *136*(1), 23-42.

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

Carlsson, F., Kataria, M., Krupnick, A., Lampi, E., Löfgren, Å., Qin, P., & Sterner, T. (2013). The truth, the whole truth, and nothing but the truth—A multiple country test of an oath script. *Journal of Economic Behavior & Organization*, *89*, 105-121.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67-80.

Duckworth, A. L., Milkman, K. L., & Laibson, D. (2018). Beyond willpower: Strategies for reducing failures of self-control. *Psychological Science in the Public Interest*, *19*(3), 102-129.

Feld, L. P., & Frey, B. S. (2018). Illegal, immoral, fattening or what?: How deterrence and responsive regulation shape tax morale. In *Size, causes and consequences of the underground economy* (pp. 15-37). Routledge.

Feldman, Y. (2017). Using behavioral ethics to curb corruption. *Behavioral Science & Policy*, *3*(2), 86-99.

Feldman, Y., van Rooij, B., & Rorie, M. (2019). Rule-breaking without Crime: Insights from Behavioral Ethics for the Study of Everyday Deviancy. *The Criminologist*, *44*(2), 8-11.

Hayes, A. F. (2015). An index and test of linear moderated mediation. *Multivariate behavioral research*, *50*(1), 1-22.HM Revenue & Customs (2019). *HMRC Digital Prompts*, 16th October 2019.

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

Hertwig, R., & Mazar, N. (2022). Toward a taxonomy and review of honesty interventions. *Current Opinion in Psychology*, 101410.

Jacquemet, N., Joule, R.V., Luchini, S. and Shogren, J.F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management, 65*(1), 110-132.

Jacquemet, N., Luchini, S., Malezieux, A. and Shogren, J.F. (2020). Who'll stop lying under oath? Empirical evidence from tax evasion games. *European Economic Review, 124*, 103369.

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

Kettle, K. L., & Häubl, G. (2011). The signature effect: Signing influences consumption-related behavior by priming self-identity. *Journal of Consumer Research*, 38(3), 474-489.

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

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

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.

Mulder, L., Jordan, J., & Rink, F. (2015). The effects of specific and general rules on ethical decisions. *Organizational Behavior and Human Decision Processes*, 126, 115-129.

Le Maux, B., & Necker, S. (2023). Honesty nudges: Effect varies with content but not with timing. *Journal of Economic Behavior & Organization*, *207*, 433-456.

Nogami, T., & Takai, J. (2008). Effects of anonymity on antisocial behavior committed by individuals. *Psychological reports*, *102*(1), 119-130.

Peer, E., & Feldman, Y. (2021). Honesty pledges for the behaviorally-based regulation of dishonesty. *Journal of European Public Policy*, *28*(5), 761-781.

Pittarello, A., Leib, M., Gordon-Hecker, T., & Shalvi, S. (2015). Justifications shape ethical blind spots. *Psychological science*, *26*(6), 794-804.

Rogers, T., & Milkman, K. L. (2016). Reminders through association. *Psychological science*, *27*(7), 973-986.

Rogers, T., Milkman, K. L., & Volpp, K. G. (2014). Commitment devices: using initiatives to change behavior. *JaMa*, *311*(20), 2065-2066.

Schlesinger, H. J. (2011). *Promises, oaths, and vows: on the psychology of promising*. Taylor & Francis.

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

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

Social and Behavioral Science Team (2015). *Annual Report*. https://github.com/gsa-oes/SBST-NSTC/blob/master/download/2015%20SBST%20Annual%20Report.pdf. Accessed 15 Feb 2023.

Swann, W. B., Jr., & Buhrmester, M. D. (2012). Self-verification: The search for coherence. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 405–424). The Guilford Press.Thaler, R. H., & Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal ofPpolitical Economy*, *112*(S1), S164-S187.

Sunstein, C. R. (2020). Sludge audits. *Behavioural Public Policy*, 1-20.

Teodorescu, K., Plonsky, O., Ayal, S., & Barkan, R. (2021). Frequency of enforcement is more important than the severity of punishment in reducing violation behaviors. *Proceedings of the National Academy of Sciences Oct 2021, 118*(42).

Wilkinson-Ryan, T., & Baron, J. (2009). Moral judgment and moral heuristics in breach of contract. *Journal of Empirical Legal Studies*, *6*(2), 405-423.

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