Journal of Experimental Social Psychology 113 (2024) 104614

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

How pledges reduce dishonesty: The role of involvement and identification☆

Eyal Peer [a,*], Nina Mazar [b], Yuval Feldman [c], Dan Ariely [d]

[a] School of Public Policy and Governance, Hebrew University of Jerusalem, Mount Scopus, Jerusalem 59000, Israel
[b] Questrom School of Business, Boston University, 595 Commonwealth Ave, Boston, MA 02215, USA
[c] Faculty of Law, Bar-Ilan University, Ramat Gan 5290002, Israel
[d] Fuqua School of Business, Duke University, 100 Fuqua Drive, Durham, NC 27708, USA

ARTICLE INFO

Keywords:
Honesty pledges
Unethical behavior
Honesty nudges

ABSTRACT

Authorities and managers often rely on individuals and businesses' self-reports and employ various forms of honesty declarations to ensure that those individuals and businesses do not over-claim payments, benefits, or other resources. While previous work has found that honesty pledges have the potential to decrease dishonesty, effects have been mixed. We argue that understanding and predicting when honesty pledges are effective has been obstructed due to variations in experimental designs and operationalizations of honesty pledges in previous research. Specifically, we focus on the role of whether and how an ex-ante honesty pledge asks individuals to identify (by ID, name, initials) and how much involvement the pledge requires from the individual (low: just reading vs. high: re-typing the text of the pledge). In four pre-registered online studies ($N > 5000$), we systematically examine these two dimensions of a pledge to find that involvement is often more effective than identification. In addition, low involvement pledges, without any identification, are mostly ineffective. Finally, we find that the effect of a high (vs. low) involvement pledge is relatively more persistent across tasks. Yet, repeating a low involvement pledge across tasks increases its effectiveness and compensates for the lower persistency across tasks. Taken together, these results contribute both to theory by comparing some of the mechanisms possibly underlying honesty pledges as well as to practice by providing guidance to managers and policymakers on how to effectively design pledges to prevent or reduce dishonesty in self-reports.

In their attempts to reduce or prevent unethical behaviors, or to ensure compliance with policies and regulations, authorities often require individuals or businesses to declare that they are or will be reporting or acting in adherence to state or organizational rules and standards. For example, witnesses swear to tell the truth in their testimony, students must state they will not cheat on their exam, employees sign contracts with honesty clauses, vendors 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, ex-ante pledges, oaths, or honor codes are the most frequently studied interventions and nudges (Hertwig & Mazar, 2022). While two recent experimental paper (Kristal et al., 2020; Le Maux & Necker, 2023) refuted previous claims and showed that merely moving the location of an honesty declaration does *not* change levels of self-reports in laboratory online and offline cheating tasks, several laboratory studies have shown honesty declarations to be effective in reducing dishonest self-reports. These later 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 tasks (Beck, Bühren, Frank, & Khachatryan, 2018; Heinicke, Rosenkranz, & Weitzel, 2019; Jacquemet, Luchini, Rosaz, & Shogren, 2019; Le Maux & Necker, 2023; Peer & Feldman, 2021). For example, Peer and Feldman (2021) found that when participants were asked to self-report their performance in an

☆ This paper has been recommended for acceptance by Paul Conway.

* Corresponding author at: Room 1742, Mexico Wing, Mount Scopus, 59000, Jerusalem, Israel.
E-mail address: eyal.peer@mail.huji.ac.il (E. Peer).

https://doi.org/10.1016/j.jesp.2024.104614
Received 19 February 2023; Received in revised form 19 February 2024; Accepted 5 March 2024
Available online 13 March 2024

online version of the matrix task (Mazar, Amir, & Ariely, 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 would report a problem as solved only if they indeed found the solution to it. Other studies have also shown that honesty pledges can encourage subsequent honest responses in preference elicitation surveys (Carlsson et al., 2013; Jacquemet, Joule, Luchini, & Shogren, 2013).

In contrast, some laboratory studies have not been successful in reducing cheating with honesty pledges (in comparison to no honesty declaration). For example, Kristal et al. (2020), (Study 1) failed to find any significant effect in a die roll task, which may have been in part due to the very low to no dishonesty observed in the control condition (without an honesty declaration), and in part due to the choice of language (i.e., content) of the honesty pledge. That pledge 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.*"), which may have hampered the pledge's ability to strengthen the compass for ethical behavior. Additionally, Cagala, Glogowsky, Rincke, and Schudy (2024) found that no-cheating declarations, that referred to the importance of ethical behavior or to possible sanctions, were ineffective in reducing cheating about the results of a random draw (for additional payoff). Furthermore, Cagala, Glogowsky, and Rincke (2021) found that asking students to sign an honesty pledge before the start of an exam ("*I hereby declare that I will not use unauthorized materials during the exam. Furthermore, I declare neither to use unauthorized aid from other participants nor to give unauthorized aid to other participants.*") significantly and substantially backfired: it at least doubled the amount of cheating compared to no honesty pledge. The authors suggest that this unexpected effect was possibly because students in the pledge condition in comparison to the no pledge condition may have believed that cheating is more prevalent. Indeed, another study found that sometimes moral reminders can increase, rather than decrease, cheating, because they provide a signal that is cheating is a foreseeable possibility (Zhao, Dong, & Yu, 2019).

A few field experiments also examined honesty pledges, though it is not always clear if the tested pledges were introduced to support existing ex-post honesty statements (i.e., honesty declarations at the end) or were introduced in a setting without any honesty declarations. For example, the Social and Behavioral Science Team in the US (Social and Behavioral Science Team, 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. Yet, a field study in Guatemala failed to find an effect of a pledge on tax reporting (Kettle, Hernandez, Sanders, Hauser, & Ruda, 2017). In this study, taxpayers were confronted with an honesty pledge in a "CAPTCHA" pop-up window (i.e., below the requirement to type the characters from the CAPTCHA picture, people saw "*Declaration: I will fill out this form honestly.*") that they needed to sign by typing their name, in order to get rid of the CAPTCHA pop-up window and proceed to complete the standard online tax return forms.. The authors speculated that the null effect may have been due to taxpayers perceiving the honesty pledge as too disconnected from the actual tax return form. They also speculated 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 tax form. Further supporting these hypotheses, is the fact that not only did the honesty pledge not have an effect, but none of the research's other behavioral interventions, which have been shown to be successful elsewhere in improving truthful self-reports, had any effects.

Together, these findings suggest that in contexts without any other honesty declarations, introducing ex-ante honesty pledges can be a useful tool to promote ethical conduct. However, it is unknown how sensitive their effectiveness may be to different operationalizations. From a practical and policy perspective, it is important to understand what aspects moderate an honesty declarations' ability to reduce self-benefiting misreporting. For this, it is critical to first distinguish so-called honesty statements versus pledges. Honesty *statements* are 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 line with most of the previous work, our research focuses only on the potential effects of such ex-ante honesty pledges in contexts without any other honesty declaration.

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 psychological mechanism is that once a person indicates to agree with a pledge, that agreement is perceived as a pre-commitment that creates a moral obligation to keep one's promise to behave ethically (e.g., Wilkinson-Ryan & 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, 2013; Swann & Buhrmester, 2012). Yet another mechanism through which honesty pledges have been proposed to operate is that they may reduce individuals' ability to morally disengage when subsequently facing an ethical dilemma (Bandura, 1989, 1990). That is, honesty pledges may make ethics and ethical conduct more salient and/or they may make it harder to disambiguate what one is expected to do, such that unethical behavior becomes harder to dismiss and/or justify (see supporting evidence from the UK HMRC context as reported by Williams & Crossfield, 2019; see also Boussalis, Feldman, & Smith, 2018; Dana, Weber, & Kuang, 2007; de Bruin, 2016; Mulder, Jordan, & Rink, 2015; Pittarello, Leib, Gordon-Hecker, & Shalvi, 2015; Shalvi, Gino, Barkan, & Ayal, 2015).

In addition to the actual text of a pledge, the degree by which it makes ethical behavior more salient may depend, among others, also on how individuals are asked to make the pledge. 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 handwrite their names or signature, etc. Thus, it appears that "not all pledges are created equal", and it is unclear what dimensions of a pledge need to be carefully designed and 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 contribute to the advancement of this goal. We do so by exploring the contribution of two specific dimensions that may vary between pledges: identification and involvement.

1. 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, increase engagement and deter people from providing false reports in fear of increased enforcements and sanctions (e.g., Gunningham, 2010; Teodorescu, Plonsky, Ayal, & Barkan, 2021). 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 consequences but also based on more internal or psychological consequences (e.g., Mazar et al., 2008). From a psychological perspective, for a pledge to be effective, it should emphasize both ethical considerations and self-perception. This way, when confronted with a temptation to misbehave, individuals will feel a significant ethical dissonance that will curb their unethical behavior (Barkan, Ayal, & Ariely, 2015).

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. 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-benefiting dishonesty (i.e., antisocial behavior and rule breaking) when their identity (e.g., 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). Together, the previous research suggests that asking for a commitment to a pledge through a signature can increase engagement depending on its operationalization (i.e., how identifiable it is). However, that by itself may not result in sufficiently high levels of engagement with the actual content of the pledge to make ethical conduct salient. To enhance the latter, we introduce the concept of "involvement" in addition to the existing concept of "identification", which we operationalize as high vs. low when requiring a person to retype a pledge vs. just reading it.

2. The current research

The empirical evidence about the effectiveness of honesty pledges (in comparison to no honesty declarations) has been mostly positive but not systematic enough to inform the implementation of pledges as means to prevent dishonesty. We contribute to a more systematic examination (with replications) of some of the factors that may make a pledge effective. 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 the 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 a pledge is more effective than merely asking for identification and whether involvement and identification have different effects in repeated instances.

To conduct this examination more systematically, we chose to build our work on the experimental design employed by Peer and Feldman (2021), using their online matrix search task and pledge on participants recruited from Prolific. This design choice allows for testing the reliability of our results but comes at the expense of generalization across different contexts (e.g., different pledge formulations, cheating tasks, recruitment platforms, modes: digital vs. paper & pencil). We conducted a series of online studies to examine different operationalizations of the two main factors of identification and involvement. A high level of identification was operationalized in varying ways, such as typing an ID (Studies 1 and 2) or signing (handwriting) one's name (Studies 2 and 3). A high level of involvement was operationalized in all studies as the requirement to re-type the text of the pledge[1] (as in Peer & Feldman, 2021) rather than just reading the text of the pledge.

Cheating task design and procedure. In the common task that was used in all studies, participants recruited from the United States on Prolific were invited to a study about problem solving in exchange for a fixed payment for completing the study and an additional bonus according to their reporting. First, similar to the design in Peer and 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 within a twelve numbers-matrix the two numbers that added up to exactly 10 (as in Mazar et al., 2008) and to do so in 20 s or less (see Fig. 1). Participants were then asked to summarize, briefly (with at least 50 characters) and in their own words, the instructions of the task. Following that, 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 s or less. Finally, participants were given additional instructions according to the condition to which they were randomly assigned. Throughout this introduction, accuracy ("exactly 10") was stressed repeatedly and was also part of the question on top of each matrix (see Fig. 1).

In the **Control** condition, participants read that they would go through X problems and earn a bonus payment 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 the bonus. 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.

In the **Self-Report** condition, which served as our baseline cheating

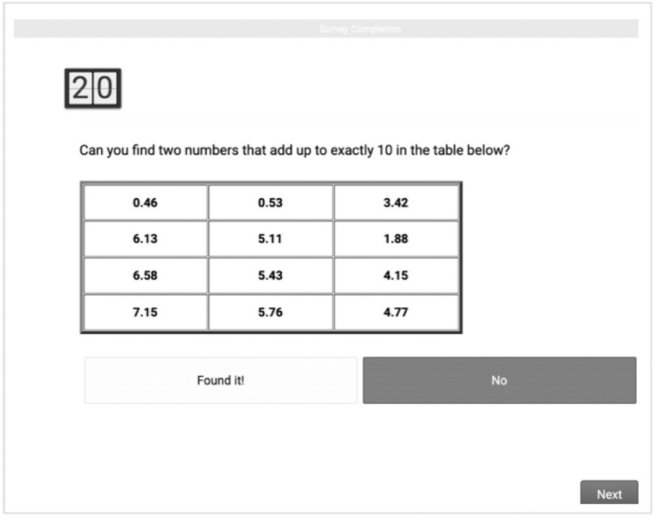

Fig. 1. Example of a problem in the Online Matrix Task.
Note: Participants had 20 s of time to actively click the "Found it!" button to gain a bonus for the problem. After 20 s, if no action had been taken, the default "No" response was recorded and the survey page automatically advanced to the next problem.

[1] The text of the pledge was presented as an image file to prevent copy-and-paste.

condition, we removed the validation step such that reporting more problems as solved resulted in higher pay. That is, participants were told that they would go through X problems and earn a bonus payment for each problem for which they indicated they found the solution (i.e., marked "Found it!"). Contrary to the Control condition, here participants were made to understand that their bonus would be determined only by the number of problems they self-reported as solved. This means that participants were led to understand that they could lie about any of the problems and report them as solved to increase their bonus.

In the **Pledge** conditions, participants were given the same instructions and design as in the Self-Report condition but were asked to first commit to a pledge before starting the matrix task. The pledge read: "*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*" (text adopted from Peer & Feldman, 2021). The manner of committing to the pledge was manipulated between-subjects, as detailed in each study.

The order of the problems was set randomly in advance and was identical for all participants across all conditions. Some of the problems were unsolvable as they did not include any combination of two numbers that added up exactly to 10. After completing the 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 were offered the opportunity to 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 a research assistant) and participants in the other conditions were paid according to the number of problems they self-reported as solved.

The main dependent variable was the number of problems that participants reported as solved (or actually solved in the Control condition[2]). The secondary dependent variable was the number of unsolvable problems that participants claimed to have found a solution for.

Overview of Studies. In Study 1 we operationalized high identification as typing an ID number (as is required in many government forms) and tested a fully crossed pledge design of high/low identification x high/low involvement. Study 1 found that the high involvement pledges reduced over-reporting significantly, whereas low-involvement pledges did not. In Study 2 we focused on testing different operationalization of high identification with requests for more personal, albeit still anonymous[3] signatures (operationalized as handwriting or typing one's first name or initials) and compared those "enhanced" forms of high identification (all with low involvement) to either high identification by typing an ID (the one used in Study 1 in combination with low involvement) or to high involvement (in combination with low identification). Study 2 found that enhanced high identification increased the effectiveness of the pledge to be comparable to high involvement, while high identification by ID was again ineffective.

After having found in Study 2 that signing one's name is relatively more effective than typing one's ID, in Study 3, we conceptually replicated Study 1 with its fully crossed design but this time with the former enhanced form of identification. Study 3 found that the combination of high involvement and high identification led to the largest reduction in dishonest reporting, while the other combinations also showed significant, albeit smaller, effects. In addition, we found that participants in the high involvement conditions were better able to recall the content of the pledge (i.e., they seemed to have paid more attention to the content of the pledge when asked to type it).

Lastly, because we found that involvement seemed more effective in reducing over-reporting than identification, in Study 4 we focused only on high vs. low involvement and examined its effect when repeating it over two consecutive tasks with a short delay (an interference task) between them. In this study we found that the effect of typing the pledge continued to last even without repeating that intervention, while that of only reading the pledge dissipated. In addition, we found that repeating the low involvement pledge helped counteract the observed decay in its effect.

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

3. Study 1 – Identification vs. involvement

The first study examined the effects of four different honesty pledges that did or did not include a request for identification and asked for either low or high involvement in making the pledge (fully-crossed design), and compared them to two conditions without any pledge (*Control* and *Self-Report*).

3.1. Method

Participants. Nine hundred and one participants 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).

Design and procedure. Four conditions manipulated the two levels of our two factors identification (low: no identification vs. high: type ID) and involvement (low: Read vs. high: Copy) in a fully crossed design. 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 resulted in higher bonus pay), the design included six between-subjects conditions. In the standard "Read" condition, participants were asked to read the text of the pledge 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 type their Prolific Participant ID to declare they agreed with the pledge. In the "Copy" condition, participants were asked to type the text of the pledge into an open text box to declare they agreed with it. In the "Copy + ID" condition, participants were asked to type both the text of the pledge and their Prolific Participant ID. Participants were paid 0.5 GBP for completing the study and an additional bonus of up to 2 GBP, that is 0.1 GBP for each problem. 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.

3.2. Results

Using G*Power software (version 3.1), a sensitivity analysis with alpha = 0.05 and power = 0.80, we found that our sample size was sufficient to detect a main effect size of $f = 0.12$. An overall ANOVA showed that the number of problems solved varied significantly between the six conditions, $F(5, 870) = 13.94$, $p < .001$, $\eta^2 = 0.08$. As can be seen in Table 1, as expected, participants' report was lowest in the Control condition (M = 5.95 problems *actually* solved, SD = 3.99), and significantly higher in the Self-Report condition (M = 9.69, SD = 5.08, $t(290.66) = 7.15$, $p < .001$, 95% CI_{diff} [5.95, 9.69], Cohen's d = 0.58),

[2] Due to a subset of unsolvable matrixes, Control condition participants could not reach 100% actually solved.

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

Table 1
Differences in self-reported performance between conditions in Study 1.

Condition	N	All problems		Unsolvable Problems	
		Mean (SD)	Test of difference from Self-Report	Mean (SD)	Test of difference from Self-Report
Control*	150	5.95 (3.99)		0.43 (0.78)	
Self-Report	155	9.69 (5.08)		0.83 (1.08)	
Read	150	9.05 (4.80)	t(302.82) = 1.13, p = .131, d = 0.13	0.77 (1.05)	t(302.97) = 0.49, p = .314, d = 0.05
Read + ID	142	10.06 (4.80)	t(294.72) = −0.64, p = .738, d = −0.07	0.83 (1.12)	t(290.44) = −0.04, p = .516, d = −0.004
Copy	140	8.77 (5.02)	t(290.65) = 1.56, p = .060, d = 0.18	0.66 (1.04)	t(291.48) = 1.36, p = .087, d = 0.16
Copy + ID	139	8.40 (4.75)	t(291.52) = 2.26, p = .012, d = 0.26	0.60 (0.94)	t(291.88) = 1.88, p = .031, d = 0.22

* Mean number of problems correctly solved.

when participants had the opportunity to over-report for higher pay.[4] That is, in the Self-Report condition, participants over-reported on average by 39% (relative to the Control condition). Further, as can be seen in Table 1, three of the four pledge conditions showed a decrease in self-reports in comparison to the Self-Report condition. Statistically analyzing the differences in report levels between each of the four pledge conditions and the Self-Report condition, however, showed that the difference was only significant for the Copy + ID condition.

Unsolvable problems. Table 1 shows that, as expected, participants in the Self-Report condition claimed more unsolvable problems as solved than participants in the Control condition, $t(281.15) = 3.65$, $p < .001$, Cohen's $d = 0.44$. In addition, compared to the Self-Report condition, only the Copy + ID condition showed a significant decrease in the number of unsolvable problems reported as solved.

4. Discussion

The results of Study 1 show that pledges can reduce people's propensity to over-report for higher pay (i.e., cheating), which is consistent with previous studies (e.g., Beck et al., 2018; Jacquemet et al., 2019; Peer & Feldman, 2021). In addition, we extend these results with a new insight about the different ways to commit to the pledge, that we had tested in this study. In particular, ensuring both a) involvement (operationalized by requiring participants to exert effort and type the text of the pledge; as in Peer & Feldman, 2021) and b) identification (operationalized by requiring participants to type their ID), may be the most effective way to curb dishonesty. Our corresponding intervention (Copy + ID condition) had on average the largest effect on curbing dishonest behavior and was the only intervention with a significant effect (comparing to the Self-Report condition). Notably, looking at the mean percent of problems reported as solved, asking to type the pledge without entering an ID (Copy) resulted in lower report levels than asking to read the pledge and enter an ID (Read + ID), though not significantly so. This latter result suggests that involvement might have played a larger role than identification in the significant effect of the Copy + ID condition.

One limitation of Study 1 is that one can claim that providing the Prolific ID is not a very strong means of identification as it may not feel as personal and related to one's self as providing one's name or initials. Moreover, previous studies showed that if awareness to the self is increased (e.g., by the message "don't be a cheater") it can reduce unethical behavior (Bryan, Adams, & Monin, 2013). Thus, in the next study, we examined other forms of identification. In particular, in addition to the Read + ID condition, we designed three "enhanced" means of identification that required signing by handwriting one's first name (Read + Sign Name) or one's initials (Read + Sign Initials), or by typing one's initials (Read + Type Initials), and compared those four high identification (but low involvement) conditions to a pledge that required no identification but high involvement (Copy).

5. Study 2 – Enhanced identification

5.1. Method

Participants. We recruited 1401 participants 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).

Design and procedure. The procedure was identical to Study 1 except that we had participants solve 10 problems[5] and they could receive a bonus of 0.2 GBP per problem (i.e., half the problems and double the bonus per problem as in Study 1). Thus, same as in Study 1, participants were paid 0.5 GBP plus a bonus of up to 2 GBP. As in Study 1, participants were randomly assigned to either a Control condition (with no option to cheat for higher bonus), a Self-Report condition, or to variations of pledge conditions. Four of the five pledge conditions asked participants to read the pledge (low involvement) and to provide an identification either in the form of handwriting their first name or initials, or in the form of typing their initials or Prolific Participant ID (Read + ID condition as in Study 1). The fifth pledge condition asked participants to copy the text of the pledge (high involvement without any identification; Copy condition as in Study 1). Two problems (matrix #2 and #8) were unsolvable.

5.2. Results

Sensitivity analysis with alpha = 0.05 and power = 0.80 showed that the sample size was sufficient to detect a small main effect size of $f = 0.10$. An overall ANOVA found statistically significant differences between the seven conditions in the percent of problems reported as solved, $F(6, 1320) = 13.42$, $p < .01$, $\eta^2 = 0.06$. Table 2 shows, as expected, that participants' reports were lowest in the Control condition (M = 2.90 problems *actually* solved, SD = 2.52), and highest in the Self-Report condition (M = 5.3, SD = 2.68, $t(328.71) = 8.59$, $p < .001$, 95% CI [1.85, 2.95], $d. = 0.95$), when participants could cheat for higher pay– a mean over-report of 45%.[6] The differences between the Self-Report condition and each of the three enhanced identification pledges, as well as the Copy condition, were statistically significant (see Table 2). In contrast, the Read + ID condition – replicating a result from Study 1 – did not significantly reduce self-reported performance levels. However, the five pledge conditions did not differ significantly from each other, $F(4, 972) = 0.28$, $p = .89$, 95% CI, effect size.

Unsolvable problems. As seen in Table 2 and similar to the results over all problems, participants in the Self-Report condition claimed more

[4] Comparing the mean of the Self-Report condition (M = 9.69, SD = 5.08) to the number of problems *claimed* as (i.e. not actually) solved in the Control condition (M = 7.44, SD = 4.25) still showed a significant reduction of 23%, $t(296.82) = 4.198$, $p < .001$, 95% CI [7.44, 9.69], Cohen's d = 0.49.

[5] 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 stated it had only 10 problems). As the materials on OSF show, the task had indeed only 10 problems.

[6] Comparing the mean of the Self-Report condition (M = 5.30, SD = 2.68) to the number of problems *claimed* as solved in the Control condition (M = 4.24, SD = 2.71) still showed a significant reduction, $t(394.46) = 3.94$, $p < .001$, 95% CI [0.53, 1.59], $d = 0.40$.

Table 2
Differences in self-reported performance between each of the pledge conditions and the Self-Report condition in Study 2.

		All problems		Unsolvable Problems	
Condition	N	Mean (SD)	Test of difference from Self-Report	Mean (SD)	Test of difference from Self-Report
Control*	149	2.90 (2.52)		0.20 (0.4)	
Self-Report	201	5.30 (2.68)		0.30 (0.46)	
Read + ID	192	4.86 (2.63)	t(390.7) = 1.64, p = .051, d = 0.17	0.27 (0.44)	t(390.99) = 0.83, p = .203, d = 0.08
Read + Sign Name	194	4.63 (2.66)	t(392.66) = 2.51, p = .006, d = 0.25	0.22 (0.41)	t(390.85) = 1.98, p = .024, d = 0.20
Read + Sign Initials	202	4.62 (2.66)	t(400.94) = 2.58, p = .005, d = 0.26	0.23 (0.42)	t(397.29) = 1.72, = .043, d = 0.17
Read + Initials	195	4.68 (2.5)	t(393.44) = 2.41, p = .008, d = 0.24	0.18 (0.38)	t(385.47) = 2.91, p = .002, d = 0.30
Copy	194	4.71 (2.57)	t(392.99) = 2.24, p = .013, d = 0.23	0.24 (0.43)	t(392.52) = 1.36, p = .086, d = 0.14

* Mean number of problems correctly solved.

unsolvable problems as solved than participants in the Control condition, t (390.62) = 2.29, p = .01, d = 0.23. In addition, compared to the Self-report condition, all three enhanced identification pledges had significantly lower levels of cheating. Same as in Study 1, the Copy condition and the Read + ID condition did not significantly reduce cheating. However, the five pledges conditions did not differ significantly from each other, F (4, 927) = 1.13, p = .34, η^2 = 0.001.

6. Discussion

Same as in Study 1, we found that asking participants to read the pledge and type their Prolific ID (Read + ID) to confirm their agreement with the pledge was ineffective at reducing overreporting and cheating (same also as in Peer & Feldman, 2021). However, we found mixed effects for the Copy condition: a significant ability to reduce over-reports over all problems solved but no significant ability to reduce cheating on the number of unsolvable problems reported as solved. For comparison, we found no significant effects in Study 1 on both dependent variables. In addition, we found that more enhanced forms of identification that involved handwriting one's name or handwriting or typing one's initials all significantly reduced overreporting consistently on both DVs. Thus, it appears that identification may be at least as likely to reduce over-reporting in comparison to involvement when it is enhanced and includes a personal piece of information (initials or name – typed or signed). This suggests that common forms, which ask individuals only to check a box or mark their agreement with a statement might not be as effective in producing accurate self-reports, and that enhancing the identification required in a pledge is important.

Building on these insights, in the next study, we aimed to replicate Study 1 (excluding the Control condition) but with an enhanced identification pledge (i.e., handwriting one's name instead of typing one's ID). That is, we aimed to test if it would replicate that a pledge that requires both high involvement and identification (Copy + Sign Name) would produce a larger effect than each of these two factors separately. In addition, we aimed to test if an enhanced identification pledge alone (Read + Sign Name) would be effective (unlike the non-enhanced identification pledge alone).

Finally, Study 3 aimed to examine potential mechanisms underlying the effectiveness of identification and/or involvement. For that purpose, we included two additional measures. First, we examined whether the effects could be the result of increased attention to the content of the pledge. In particular, we predicted that high involvement (asking participants to type the pledge) would result in higher attention to the content of the pledge, manifested as better recall of its content, which in turn would drive more accurate reporting. As another potential mediator, we examined how identification and/or involvement elicited concerns of being sanctioned for cheating. To decrease noise around the examination of these two mechanisms, we decided to focus our study on respondents who would pass a couple of attention check questions before the start of our study and pre-registered so.

7. Study 3 – Enhanced identification vs. involvement

7.1. Method

Participants. The sample included 1498 participants (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 was achieved after excluding participants from an original sample of 1993 who said they completed a similar study in the past (N = 260), did not follow instructions of the pledge (N = 3), and that failed a general attention-check question (placed right at the beginning of the study, after the consent form but before the task instructions; N = 232).

Design and procedure. Same as Study 2, Study 3 had only 10 problems, and participants received 0.5 GBP for completing the study and a bonus of up to 2 GBP according to their performance (0.2 GBP per problem). Unlike the previous two studies, the time limit to solve each problem was reduced to 10 s. Further, half of the problems (the even numbered ones) were unsolvable. This study had five conditions (this time we omitted the Control condition) that included a Self-Report condition and four pledge conditions: Read, Read + Sign Name, Copy, Copy + Sign Name. Signing was operationalized by handwriting one's first name.

Additionally, we explored two potential mediators at the end of the study: First, we measured attention to the pledge by asking participants to select, out of six quotes (three correct and three incorrect), the ones that they thought had appeared in the text of the pledge (aided recall procedure adopted from Shu, Gino, & Bazerman, 2011). The pledge attention score was calculated as the difference between the number of correctly selected quotes versus incorrectly selected quotes, yielding a score ranging from (−3) to (+3). Second, we explored perceived fear of being sanctioned by the researcher for cheating. To assess that fear we asked participants to indicate their agreement (from 1: strongly disagree to 5: strongly agree) with three statements: "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", and "I was sure that my submission will be approved, regardless of how many problems I report as solved" (reverse coded).

7.2. Results

Sensitivity analysis with alpha = 0.05 and power = 0.80 showed that the final sample size was sufficient to detect a small main effect size of f = 0.09. An overall ANOVA found significant differences between the five conditions in the number of problems reported as solved, both for all problems, F (4, 1493) = 9.34, p < .01, 95% CI, η^2 = 0.02, and for the unsolvable problems, F (4, 1493) = 6.84, p < .01, η^2 = 0.01. As can be seen in Table 3, unlike in Study 1, this time all four types of pledges significantly reduced reported performance in comparison to the Self-Report condition (all pairwise p < .001). Compared to the Self-Report

Table 3
Differences in self-reported performance between each of the pledge conditions and the Self-Report condition in Study 3.

Condition	N	All problems		Unsolvable Problems	
		Mean (SD)	Test of difference from Self-Report	Mean (SD)	Test of difference from Self-Report
Self-Report	301	3.70 (2.76)		1.70 (1.63)	
Read	301	2.84 (2.57)	t(597.02) = 3.91, $p < .001$, $d = 0.32$	1.33 (1.53)	t(597.7) = 2.95, $p = .002$, $d = 0.24$
Read + Sign Name	300	2.96 (2.68)	t(598.59) = 3.32, $p < .001$, $d = 0.27$	1.27 (1.54)	t(597.45) = 3.39, $p < .001$, $d = 0.27$
Copy	299	2.70 (2.40)	t(587.66) = 4.73, $p < .001$, $d = 0.39$	1.21 (1.40)	t(586.54) = 3.96, $p < .001$, $d = 0.32$
Copy + Sign Name	297	2.53 (2.38)	t(585.43) = 5.58, $p < .001$, $d = 0.46$	1.12 (1.33)	t(576.92) = 4.83, < 0.001, $d = 0.40$

condition, asking participants to only read the pledge reduced reported performance by 23% and asking to read and handwrite one's name reduced it by 20%. Asking participants to type the pledge reduced reports by 27% and asking to both type the pledge and handwrite one's name reduced reports the most: by 32%. However, the differences between the four pledges conditions were not statistically significant, F (3, 1193) = 1.67, $p = .17$, $\eta^2 = 0.001$.

Exploratory 2 × 2 ANOVA Over All Problems. Deviating from or preregistration plan, we explored the differences between the four pledge conditions using a 2 × 2 ANOVA. This analysis revealed a significant main effect for involvement, F (1, 1194) = 20.02, $p < .001$, $\eta^2 = 0.02$. That is, high involvement pledges (Copy and Copy + Sign Name, M = 2.61, SD = 2.39) resulted in significantly lower reporting than low involvement pledges (Read and Read + Sign Name, M = 3.17, SD = 2.70). There was also a significant main effect of identification (Read + Sign and Copy + Sign vs. Read and Copy), F (1, 1194) = 12.72, $p < .001$, $\eta^2 = 0.01$, and a significant interaction F (1, 1194) = 35.31, $p = .01$, $\eta^2 = 0.001$.

Exploratory analysis without exclusions. Deviating from our preregistration plan, we re-analyzed the study without excluding participants (i.e., full sample of $N = 1993$) to explore the robustness of our findings. We again found significant differences between conditions in the number of problems reported as solved, both for all problems, F (4, 1988) = 12.69, $p < .001$, $\eta^2 = 0.02$, and for the unsolvable problems, F (4, 1988) = 10.94, $p < .001$, $\eta^2 = 0.02$. For the 2 × 2 ANOVA, we again found significant main effects for involvement (F (1, 1989) = 24.23, $p < .01$, $\eta^2 = 0.02$), identification (F (1, 1989) = 18.49, $p < .01$, $\eta^2 = 0.01$), and their interaction (F (1, 1989) = 6.93, $p < .01$, $\eta^2 = 0.001$).

Mediator 1: Attention to the Content of the Pledge. We examined attention to the pledge's content through aided recall, with a score ranging from (−3) to (+3) (as explained above). Recall was higher in the Copy (M = 2.22, SD = 0.90) and Copy + Sign conditions (M = 1.99, SD = 0.97) than in the Read (M = 0.93, SD = 1.12) and Read + Sign conditions (M = 0.73, SD = 1.04). An 2 × 2 ANOVA revealed significant main effects for involvement (F (1, 1193) = 477.36, $p < .001$, $\eta^2 = 0.29$) and identification (F (1, 1193) = 13.94, $p < .001$, $\eta^2 = 0.01$), but not for their interaction (F (1, 1193) = 0.08, $p = .77$, $\eta^2 = 0.001$). The effect size of the differences between conditions was high for involvement (Cohen's $d = 1.26$), and lower for identification (Cohen's $d = 0.18$). Contrary to our expectations, the mediation analysis did not show a statistically significant indirect effect of involvement on self-reports through level of recall, b = −0.13, 95% CI [−0.30, 0.06]). We also tested another mediation path by which recall score was the dependent variable and the percent of problems reported was the mediator, and again did not find a significant mediation effect, b = 0.005, 95% CI [−0.002, 0.015]).

Mediator 2: Fear of Sanctions. Due to low Cronbach's alpha (0.52) for our three items, we did not average them to one fear of sanctions-score. Instead, we analyzed for each item separately whether there was a significant difference between the pledge conditions and found that there was none (fear of misreporting: F (3, 1193) = 0.2, $p = .897$, $\eta^2 = 0.001$; fear of reporting too many problems: F (3, 1193) = 0.95, $p = .416$, $\eta^2 = 0.001$; no fear: F (3, 1193) = 0.91, $p = .434$, $\eta^2 = 0.001$). Thus, we did not proceed to running mediation analyses.

7.3. Discussion

The results of this study show that the combination of high involvement and high identification (asking to type the pledge and handwrite one's name) resulted in the largest reduction of dishonest reporting from the Self-Report condition, though the effectiveness of this pledge (Copy + Sign Name) was not significantly different from the effects of the other pledges. In addition, the study did not find empirical support attention to the content of the pledge (measured by aided recall) mediating the effect of the high-involvement pledges on self-reporting.

Summary of Studies 1–3. Because some of our pledges were tested in more than one study, we can look for patterns of consistency of their effects (in comparison to the Self-Report condition). Doing so, we found the following for involvement: The pledge that required only high involvement (Copy) had consistently significant effects on reducing self-reported performance levels. That is, the significant effect replicated across the three studies in which this high involvement + low identification pledge was used, with effect sizes ranging from $d = 0.18$ (Study 1) to $d = 0.23$ (Study 2) to $d = 0.39$ (Study 3). In contrast, the pledge that required only low involvement (Read) had inconsistent effects across the two studies where it was used. That is, the mean difference to the Self-Report condition of that low involvement + low identification pledge was non-significant in Study 1 ($d = 0.13$) but significant in Study 3 ($d = 0.32$).

We've found the following for identification: First, the pledge that required only ID identification (Read + ID) had consistently non-significant effects. That is, the non-significant effect replicated in both studies in which that low involvement + high identification pledge was used (Study 1: $d = -0.07$; Study 2: $d = 0.17$). Second, the enhanced identification pledge that required only handwriting one's name (Read + Sign Name) had consistently significant effects. That is, the significant effect on reducing self-reports replicated in both studies in which that low involvement + enhanced high identification pledge was used, and effect sizes ranged from $d = 0.25$ (Study 2) to $d = 0.27$ (Study 3). Finally, when combining involvement with identification, regardless of the type (not enhanced: typing one's ID and enhanced: handwriting one's name), we found consistently significant effects of that combination on reducing self-reports, with effect sizes ranging from $d = 0.26$ (Study 1 with Copy + ID) to $d = 0.46$ (Study 3 with Copy + Sign Name). Together, it appears that the effect of high involvement on self-reports is significant and consistent while the effect of high identification is significant and consistent only when identification is "enhanced".

8. Study 4 – Repeating pledges

The three studies thus far examined immediate effects of pledges right before a task in which participants faced a series of opportunities to cheat (i.e., for each of the problems in the matrix task). However, it is unclear to what extent effects of pledges persist across tasks, and to what extent people may habituate to them or show consistent or even stronger responses across repeated exposures (i.e., if individuals need to express their agreement with a pledge several times; Robitaille, House, & Mazar, 2021). These questions are 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). A recent study examined this persistency question with participants encountering a pledge similar to our Read condition ("…by continuing, you declare that you will honestly report your guesses.") either at the 1st (i.e. the start) or the 6th (i.e. the middle) of 10 trials in a wheel of fortune task, in each of which participants could cheat (Le Maux & Necker, 2023). While the effect of that study's pledge at either location did not deteriorate over the subsequent trials, it is not clear, to what extent it may persist after the task. In addition, like many other nudges, pledges 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; Bandura, Barbaranelli, Caprara, & Pastorelli, 1996), presents a major challenge to the use and effectiveness of pledges in real-life settings.

Thus, to contribute to answering the questions of persistency across tasks and consistency across repetitions, participants in our last Study 4 completed two matrix tasks, with a short delay between them, and half of the participants repeated the pledge after the delay (i.e., before the second matrix task). Because our previous studies found the effect of involvement to be more consistent than that of identification, in Study 4 we tested only two pledges that varied in their level of involvement (i.e., we kept identification low across both pledges): low involvement pledge (Read) and high involvement pledge (Copy).

8.1. Method

Participants We recruited 1505 participants (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 participation for those that had completed a similar study on our Prolific account. None of the participants were excluded from the final sample. Participants were paid 1 GBP plus a bonus of up to 3 GBP.

Design and procedure. First, participants participated in a matrix task with 15 problems 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 related attention check questions.[7] Then, participants were asked to participate in a second matrix task with another set of 15 problems for an additional bonus of 0.1 GBP each. Finally, after completing both matrix tasks, we asked the same mediator question used in Study 3 to assess attention to the pledge content.

Participants were randomly assigned to one of five conditions: one Self-Report and four pledge conditions (Copy once, Copy repeated, Read, Read repeated). In the two Copy conditions, participants were asked either once (before the onset of the first set of matrix task) or repeated (before the onset of each of the two matrix tasks) to consent to a similar pledge as used in the previous studies by typing it. In the two Read conditions, participants were asked either once (before the onset of the first matrix task) or repeated (before the onset of each of the two matrix tasks) to mark a checkbox that they agreed with the presented pledge.

8.2. Results and discussion

Sensitivity analysis with alpha = 0.05 and power = 0.80 showed that the sample size was sufficient to detect a main effect size of $f = 0.09$ between the five conditions. Fig. 2 and Table 4 display the means and errors in each condition for each of the two matrix tasks.

Matrix Task 1. An ANOVA found significant differences between the number of problems reported as solved between the conditions, $F(4, 1500) = 11.49$, $p < .001$, $\eta^2 = 0.03$. Follow-up pairwise comparisons revealed significantly lower reports each of the pledge conditions than in the Self-Report condition (all $ps <= 0.03$). Because the experimental procedure was identical in the "Once" and the "Repeated" conditions in Matrix Task 1, we collapsed the two Read condition and we collapsed the two Copy conditions, and compared their collapsed conditions to the Self-report condition. Doing so, we found significant differences between the Read and Self-report conditions ($t(557.3) = 3.47$, $p < .001$, 95% CI_{diff} [0.44, 1.59], d = 0.29) as well as between the Copy and Self-report conditions ($t(505.71) = 6.30$, $p < .001$, 95% CI_{diff} [1.23, 2.34], d = 0.56). Finally, we found that on average the Copy condition (M = 4.23, SD = 3.4) was significantly more effective than the Read condition (M = 5.01, SD = 3.88) in reducing reports, $t(1184.5) = 3.66$, $p < .001$, 95% CI_{diff} [0.36, 1.18], d = 0.21.

Matrix Task 2. We again found a significant difference between all five conditions, $F(4, 1500) = 11.93$, $p < .001$, $\eta^2 = 0.03$. Compared to the Self-report condition, participants in the Copy conditions reduced over-reports significantly, no matter if the pledge was repeated or not, (both $ps < 0.001$). However, compared to the Self-report condition, participants in the Read conditions reduced over-reports significantly only when the pledge was repeated, $p < .001$, but not when it was made once, $p = .18$. All remaining pairwise comparisons were significant, (all $ps < 0.02$), except for the pairwise comparison of the difference between the Copy Once and the Copy Repeated conditions, $p = .25$.

Matrix Tasks 1 and 2. To examine the effects of the different pledges across the two separate matrix tasks, we conducted another analysis, that was not pre-registered. Specifically, we explored the percent of participants that reported a given problem as solved, over each of the sequential problems and across each matrix tasks. As can be seen in Fig. 3, when no pledge was administered (Self-Report condition), the percentage increased slightly over problems and tasks, but this trend was not statistically significant, $b = 0.14$, SE = 0.17, $p = .396$, 95% CI (−0.17, 0.46). In the Copy-Once condition, the percentage was considerably smaller significantly increase over problems, $b = 0.49$, SE = 0.16, $p = .002$, 95% CI (0.19, 0.81]. However, there was no significant interaction between problem number (i.e. trial number) and repeating the pledge in the Copy condition, $b = -0.12$, SE = 0.22, $p = .578$, 95% CI (−0.56, 0.31), suggesting the trend over problems was similar in the Matrix Task 2 whether the pledge was repeated or not. In the Read-Once condition, the percentage in the Matrix Task 1 started higher than in the Copy-Once condition. However, there was a significant interaction between problem number and repeating the Read pledge, $b = -0.49$, SE = 0.23, $p = .033$, 95% CI (−0.94, −0.04), showing that after the Read pledge was repeated before the Matrix Task 2, the percentage of participants that reported a problem as solved decreased, and when the pledge was not repeated, the percentage increased.

Mediator: Attention to the Content of the Pledge. As in Study 3, we again found higher attention rates (aided recall score) among those who had to Copy the pledge, once (M = 1.71, SD = 1.01) or repeated (M = 2.32, SD = 0.88) than those who only had to Read the pledge, once (M = 0.52, SD = 1.10) or repeated (M = 0.95, SD = 1.04). ANOVA on recall levels showed significant effects for level of involvement ($F(1, 1195) = 498.8$, $p < .001$, $\eta^2 = 0.29$) and for whether the pledge was repeated ($F(1, 1195) = 79.8$, $p < .001$, $\eta^2 = 0.06$), but not for the interaction of these two factors ($F(1, 1195) = 2.261$, $p = .133$, $\eta^2 < 0.01$), confirming that copying the pledge and repeating it leads to highest levels of recall of 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 to the content of the pledge (recall)

[7] 14% of the participants answered the attention check questions for the video incorrectly. However, we did not find any significant interaction between performance on these questions and the pledge conditions either on Self-Report in the second task or overall, $ps > 0.06$, and thus we did not consider this attention check-variable in our analyses.

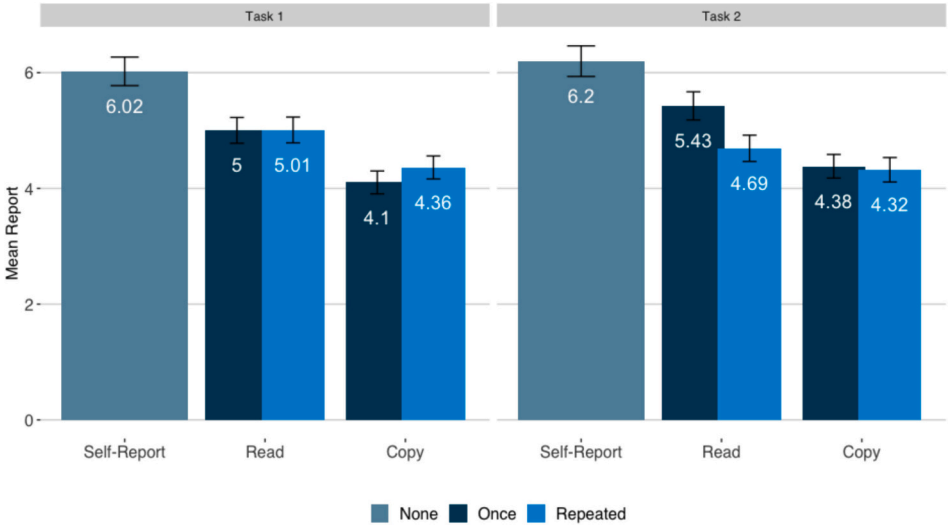

Fig. 2. Mean number of problems reported as solved, between conditions and matrix tasks in Study 4.
* Error bars show one standard error above/below the mean.

Table 4
Mean of self-reported performance in each condition and matrix task in Study 4.

		All problems Mean (SD)		Unsolvable Problems Mean (SD)	
Condition	N	Task 1	Task 2	Task 1	Task 2
Self-Report	301	6.02 (4.32)	6.20 (4.62)	2.16 (2.36)	2.16 (2.50)
Read Once	301	5.00 (3.89)	5.43 (4.26)	1.58 (2.05)	1.73 (2.29)
Read Repeat	300	5.01 (3.88)	4.69 (3.95)	1.70 (2.05)	1.43 (2.13)
Copy Once	299	4.10 (3.36)	4.38 (3.48)	1.18 (1.79)	1.21 (1.92)
Copy Repeat	297	4.36 (3.45)	4.32 (3.66)	1.32 (1.80)	1.13 (1.88)

mediated the effect of involvement, and whether repeating the pledge had an effect on that mediation, we conducted a moderated mediation analysis (Hayes, 2015; Model 7) with involvement (high: Copy vs. low: Read) as the independent variable, recall level as the mediator, whether the pledge was repeated as the moderator, and the total number of problems reported as solved (across both tasks) as the dependent variable (reports). We found that copying the pledge had a significant effect on reducing reports (b = −1.65, SE = 0.49, $p < .01$, 95% CI [−2.60, −0.70]. However, the indirect effect of recall on reports was not significant, whether the pledge was repeated (b = 0.18, SE = 0.27, 95% CI [−0.33, 0.71]) or not (b = 0.16, SE = 0.23, 95% CI [−0.29, 0.62]. The index for the moderated mediation (difference between the above indirect effects) was also not significant, b = 0.02, SE = 0.04, 95% CI [−0.05, 0.13].

To summarize, the results of this study showed that the effects of our high involvement pledge (Copy) on reducing self-reports was stronger than that of our low involvement pledge (Read), as was found in our previous studies. Extending our previous findings, we found that repeating the pledge contributed to increasing its effectiveness only when the pledge was initially of low involvement (Read). In addition, that pledge's effectiveness did not persist without repetition; in fact, it decreased. When the pledge was of high involvement (Copy), its effect persisted (i.e., the pledge continued to be effective) when not repeated and was consistent when repeated. Attention to the content of the pledge, measured by aided recall, was considerably higher when asked to copy the pledge, but this effect –as in Study 3 – did not mediate the

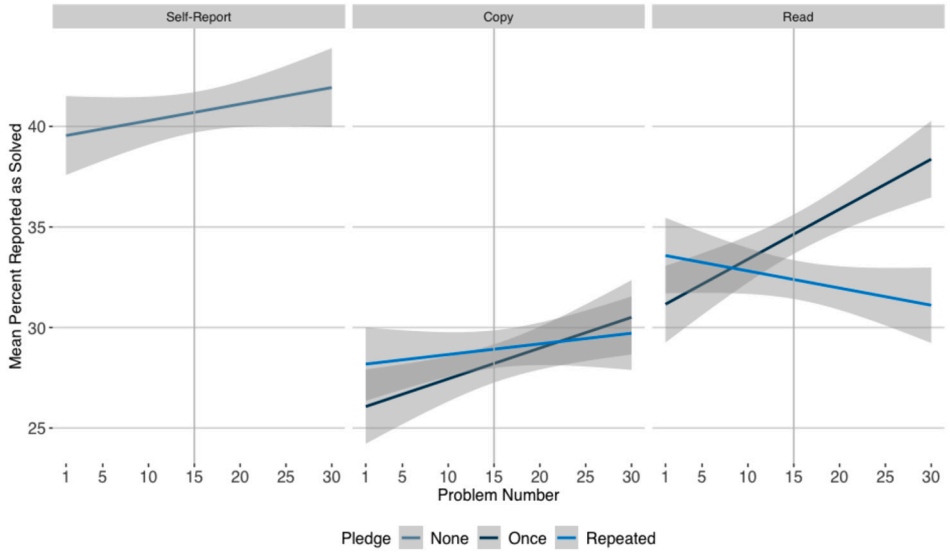

Fig. 3. Percentage of participants reporting a problem as solved over reach problem and tasks and between conditions in Study 4.
* The vertical grey line separates the problems from matrix tasks 1 (Problem #1-#15) and 2 (Problem #16-#30).

higher effectiveness of the Copy pledge on reducing reports.

9. Summary of results over all studies

Table 5 summarizes the effect sizes of all pledges used across our four studies, using Cohen's d measures of the mean difference between each pledge condition and the respective Self-Report conditions. We classified the identification conditions Sign Name, Sign Initials, and Type Initials as "Enhanced" identification. As can be seen, the effects elicited by the Copy condition in Study 4, and the Copy + Sign Name condition in Study 3 were the largest. Among all the conditions that relied on high levels of involvement, the effect was not significant only once ($d = 0.18$ in Study 1) out of the six instances (16.67% of the time), and the significant effects ranged in size from small (d = 0.23) to medium (d = 0.49). Among all the conditions that relied on low levels of involvement, the effects were not significant in three out of the nine instances (33% of the time), and the significant effects were all relatively small in size ($d =$ 0.24–0.32).

10. General discussion

Honesty pledges attempt to change the level of ethicality in peoples' behavior (e.g., de Bruin, 2016). However, previous research on honesty pledges did not systematically determine what are the factors that make pledges effective and to what extent. Understanding the relevant factors 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 practical perspective, understanding why pledges work can inform policy on the "how" to better design and implement them, and this insight 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. First, we document reliable effects of pledges that *replicated* across four *pre-registered* studies with *large sample sizes*, three aspects that are still underrepresented in the existing body of work on pledges. Second, our findings show that honesty pledges can substantially reduce unethical behavior by as much as 15–30% 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). Third, our findings also show to what extent the two factors involvement and identification matter in the design of effective pledges. In that regard we found, for example, that when pledges are minimally implemented – as by asking people to only mark a checkbox after presumably having read it – their effects are smaller and often not significant.

Together, our findings provide insights on when honesty pledges could have stronger or weaker effects on reducing unethical behavior. While signatures (i.e., identification) 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 engage with the pledge. Notable is the finding of Study 3, which directly contrasted involvement and identification. This insight is especially important in contexts where getting individuals' 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 systematic attention to such design features. 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.

Limitations. Our studies have various limitations, particularly with regards to external validity that should be addressed in future research to further explore the effectiveness of honesty pledges in the real world. For example, our experiment could not include (very) high temptations for dishonest reports. Thus, we cannot generalize our findings to situations where the potential gains are considerably larger, such as in the context of large-scale frauds. Instead, our studies may be more informative in terms of how honesty pledges may curb dishonesty in everyday lives; that is, in common low-stakes situations. Similarly, our study cannot speak to interactions with negative incentives; that is, deterrents such as sanctions, as dishonest reports could not be identified and had no consequences in our studies. More generally, we do not know to what extent the online matrix task captures which real-world scenarios where people are facing temptations for dishonest self-reports.

Second, while we discovered that involvement with the pledges had an overall stronger effect on reducing self-reports than identification, this may be a result of the particular operationalizations that we used and not a general takeaway. It would be useful to test other practical operationalizations. For example, one may increase involvement by asking pledgers to read the pledge aloud or formulate the content of the pledge themselves. Similarly, one may increase identification by asking for a selfie-photo, fingerprints, or other highly identifying information. Thus, future research might want to examine the generalizability of our work's takeaway with regards to the relative importance of each of these two components for a pledge's effectiveness.

Third, additional research may want to further examine the potential mechanisms underlying high involvement and/or high identification pledges. While our research suggested that high involvement pledges lead to higher attention to the content of the pledge (manifested in higher aided recall rates), we did not find evidence that attention mediated the effect of the pledges. This may be in part a function of how we measured attention. Similarly, the fact that we didn't find any difference between pledges when it comes to "fear of sanctions," that could be due to the specific operationalization that we chose. For example, we measured the construct only after and not before the cheating task. Thus, what we may have captured is participants' post-hoc rationalizations rather than actual differences in "fear of sanctions" triggered by the different types of pledges. Another possibility is that if we had not only focused on the external costs one may expect (i.e., their submission being rejected) but also on real moral or social costs, that then we may have seen differences for "fear of sanctions." Future research may also want to examine to what extent our relatively narrow definitions of the two components involvement and identification relate to the formation of both the private (internalization) and public (symbolization; see also Paulhus, 1991) moral identity dimensions, as put forward by Aquino and Reed (2002).

Table 5
Overview of effect sizes of pledges in Studies 1–4 (Cohen's *d* of the mean difference between each pledge and the respective self-report condition).

Pledge Classification	Level of Involvement	Type of Identification	Study 1	Study 2	Study 3	Study 4
Read	Low	None	0.13		**0.32**	**0.22 / 0.31****
Read + ID	Low	Type ID	−0.07	0.17		
Read + Enhanced	Low	Sign Name, Sign or Type Initials		**0.24–0.26***	**0.27**	
Copy	High	None	0.18	**0.23**	**0.39**	**0.49 / 0.46****
Copy + ID	High	Type ID	**0.26**			
Copy + Enhanced	High	Sign Name			**0.46**	

* Read + Initials condition: $d = 0.24$, Read + Sign Name condition: $d = 0.25$, Read + Sign Initials condition: $d = 0.26$. ** Effect size for the repeated pledge. Effects significant at $p < .05$ are in bold.

Lastly, it could be interesting to examine the actual content of the pledge and to what extent variations in it are important for its effectiveness to reduce dishonest self-reports. For example, it could be argued that the effect of our pledges stem only from the fact that the pledges spelled out the instructions for the task to participants, asking them to prioritize accuracy over personal gain and not from agreeing with the pledge. That is, our pledge manipulation included both a promise and a direct repetition and elaboration of what is considered a solution to a problem. Thus, we do not know to what extent each of these two contributed to the effectiveness of our pledges. However, because participants in the non-pledge conditions also received detailed instructions, with a request to describe those instructions in an open-text question before going through an actual practice round, and we did find significant differences between our pledge conditions and the Self-Report conditions, we posit that this explanation is not very likely. Yet, future research may want to examine these potential directions and also replicate and expand our findings to other pledge statements that could vary in framing (positive vs. negative), and/or the specificity of the behavior one is pledging to adhere to (e.g., Mulder et al., 2015) found specific rules to work better at inducing ethical decisions than general rules). The scope of future research, which can be far reaching and could not have been covered in this paper, should be expanded, and examined (including across cheating tasks and more diverse populations), to further understand the breath and boundaries of the effectiveness of honesty pledges.

From a policy perspective, our findings suggest that managers and policymakers should explore methods to enhance individuals' engagement with pledges, affidavits, oaths, contracts, or other pre-commitment certifications aimed at preventing dishonesty and ensuring compliance. Honesty pledges that succeed particularly in engaging individuals could function as a "soft" form of 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's ability for self-control (Duckworth, Milkman, & Laibson, 2018; Rogers & Milkman, 2016; Rogers, Milkman, & Volpp, 2014). Thus, if designed properly, ex-ante 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 monitoring and enforcement resources currently allocated for lengthy and costly checks and inspections (that also increase the time citizens and businesses must wait for responses) and instead focus their attention 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 and trustworthiness, may gradually improve and re-build trust between state regulators and the citizens and businesses they serve, as well as between managers and their employees or between leaders and their teams. This could, eventually, enhance social capital which can result in improved efficiency, welfare and incentivized growth in society Putnam (2001).

CRediT authorship contribution statement

Eyal Peer: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Nina Mazar:** Conceptualization, Investigation, Methodology, Supervision, Validation, Writing – review & editing. **Yuval Feldman:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Validation, Writing – review & editing. **Dan Ariely:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

We declare no conflict of interest in the manuscript "How pledges reduce dishonesty: The role of identification and involvement".

Data availability

All data are shared at https://osf.io/hda7b/?view_only=4ce2fa3bf29c4df1adb532aa1faf3dd6

Acknowledgments

The authors thank Rotem Levin for her research assistance in conducting the studies, and the financial support from the Israel Science Foundation (Award #385/20 to Eyal Peer and Yuval Feldman) and the European Research Council (Grant number: 101054656/VCOMP to Yuval Feldman). We also thank the editors and reviewers for their valuable suggestions.

References

Aquino, K., & Reed, A. (2002). The self-importance of moral identity. *Journal of Personality and Social Psychology, 83*, 1423–1440.
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.
Barkan, R., Ayal, S., & Ariely, D. (2015). Ethical dissonance, justifications, and moral behavior. *Current Opinion in Psychology, 6*(Dec), 157–161.
Beck, T., Bühren, C., Frank, B., & 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.
de Bruin, B. (2016). Pledging integrity: Oaths as forms of business ethics management. *Journal of Business Ethics, 136*(1), 23–42.
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. Revista de Economia, 2*(1), 671–698.
Cagala, T., Glogowsky, U., & Rincke, J. (2021). Detecting and preventing cheating in exams: Evidence from a field experiment. *Journal of Human Resources, 59*(1), 210–241.
Cagala, T., Glogowsky, U., Rincke, J., & Schudy, S. (2024). Commitment requests do not affect truth-telling in laboratory and online experiments. *Games and Economic Behavior, 143*, 179–190.
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.
Gunningham, N. (2010). Enforcement and compliance strategies. In R. Baldwin, M. Cave, & M. Lodge (Eds.), *The Oxford handbook of regulation* (pp. 120–145).

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., & Shogren, J. F. (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management, 65*(1), 110–132.
Jacquemet, N., Luchini, S., Rosaz, J., & 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., & 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.
Le Maux, B., & Necker, S. (2023). Honesty nudges: Effect varies with content but not with timing. *Journal of Economic Behavior & Organization, 207*, 433–456.
Mazar, N., Amir, O., & 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.
Nogami, T., & Takai, J. (2008). Effects of anonymity on antisocial behavior committed by individuals. *Psychological Reports, 102*(1), 119–130.
Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.
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.
Putnam, R. (2001). Social capital: Measurement and consequences. *Canadian Journal of Policy Research, 2*(1), 41–51.
Robitaille, N., House, J., & Mazar, N. (2021). Effectiveness of planning prompts on organizations' likelihood to file their overdue taxes: A multi-wave field experiment. *Management Science, 67*(7), 4327–4340.
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.
Sunstein, C. R. (2020). Sludge audits. *Behavioural Public Policy*, 1–20.
Swann, W. B., & 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.
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, 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.
Williams, H., & Crossfield, J. (2019). Upfront honesty declaration. HMRC report 603. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/995378/HMRC_research_report_603_upfront_honesty_declaration.pdf.
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*, Article 103895.