# I Solemnly Swear I'm Up To Good: A Megastudy Investigating the Effectiveness of Honesty Oaths on Curbing Dishonesty

Janis Zickfeld, Karolina Scigala, Christian Elbaek, John Michael, Mathilde Tønning Tønnesen, Gabriel Levy, Shahar Ayal, Isabel Thielmann, Laila Nockur, Eyal Peer, et al.

# I Solemnly Swear I'm Up To Good: A Megastudy Investigating the Effectiveness of Honesty Oaths on Curbing Dishonesty

Janis H. Zickfeld[1†], Karolina A. Ścigała[1], Christian T. Elbæk[1], John Michael[2], Mathilde H. Tønnesen[1], Gabriel Levy[3], Shahar Ayal[4], Isabel Thielmann[5], Laila Nockur[1], Eyal Peer[6], Valerio Capraro[7], Rachel Barkan[8], Simen Bø[9], Štěpán Bahník[10], Daniele Nosenzo[1], Ralph Hertwig[11], Nina Mazar[12], Alexa Weiss[13], Ann-Kathrin Koessler[14], Ronit Montal-Rosenberg[15], Sebastian Hafenbrädl[16], Yngwie A. Nielsen[1], Patricia Kanngiesser[17], Simon Schindler[18], Philipp Gerlach[19], Nils Köbis[11,20], Nicolas Jacquemet[21], Marek Vranka[22], Dan Ariely[23], Jareef Bin Martuza[9], Yuval Feldman[24], Michał Białek[25], Jan K. Woike[17], Zoe Rahwan[11], Alicia Seidl[5,26], Eileen Chou[27], Agne Kajackaite[2], Simeon Schudy[28,29], Ulrich Glogowsky[30,29], Anna Z. Czarna[31], Stefan Pfattheicher[1], Panagiotis Mitkidis[1]

[1]Aarhus University, [2]University of Milan, [3]NTNU Trondheim, [4]Reichman University, [5]Max Planck Institute for the Study of Crime, Security and Law, [6]Hebrew University of Jerusalem, [7]University of Milan Bicocca, [8]Ben-Gurion University, [9]Department of Strategy and Management, Norwegian School of Economics, [10]Prague University of Economics and Business, [11]Max Planck Institute for Human Development, [12]Boston University, [13]Bielefeld University, [14]Leibniz University Hannover, [15]The Hebrew University of Jerusalem, [16]IESE Business School, [17]University of Plymouth, [18]Federal University of Applied Administrative Services, Berlin, [19]Fresenius University, Hamburg, [20]University of Duisburg-Essen, [21]Paris School of Economics \& Université Paris 1 Panthéon-Sorbonne, [22]Charles University, [23]Duke University, [24]Bar Ilan University, [25]University of Wrocław, [26]University of Kaiserslautern-Landau, [27]University of Virginia, [28]Ulm University, [29]CESifo, [30]Johannes Kepler University Linz, [31]Jagiellonian University

†Correspondence should be addressed to Janis H. Zickfeld; E-mail: jhzickfeld@gmail.com.

Note. Authorship order for the first seven and last two authors set a-priori. Authorship order for the remaining authors determined randomly.

**Pre-print before peer review.** Scientific articles usually go through a peer-review process. This means that independent researchers evaluate the quality of the work, provide suggestions, and speak for or against the publication. Please note that the present article has not (yet) undergone this standard procedure for scientific publications.

## Author Contributions

**Conceptualization:** Janis H. Zickfeld and Panagiotis Mitkidis.

**Data curation:** Janis H. Zickfeld.

**Formal analysis:** Janis H. Zickfeld.

**Funding acquisition:** John Michael, Dan Ariely, Isabel Thielmann, Sebastian Hafenbrädl, Ann-Kathrin Koessler, Stefan Pfattheicher, and Panagiotis Mitkidis.

**Investigation:** Janis H. Zickfeld.

**Methodology:** Janis H. Zickfeld, Karolina A. Ścigała, Christian T. Elbæk, John Michael, Mathilde H. Tønnesen, Gabriel Levy, Shahar Ayal, Eyal Peer, Eileen Chou, Ulrich Glogowsky, Simeon Schudy, Yuval Feldman, Patricia Kanngiesser, Nina Mazar, Simon Schindler, Alexa Weiss, Daniele Nosenzo, Valerio Capraro, Ralph Hertwig, Rachel Barkan, Dan Ariely, Ronit Montal-Rosenberg, Isabel Thielmann, Jareef Bin Martuza, Philipp Gerlach, Štěpán Bahník, Michał Białek, Jan K. Woike, Laila Nockur, Simen Bø, Nicolas Jacquemet, Nils Köbis, Sebastian Hafenbrädl, Zoe Rahwan, Agne Kajackaite, Yngwie A. Nielsen, Ann-Kathrin Koessler, Marek Vranka, Anna Z. Czarna, Alicia Seidl, Stefan Pfattheicher, and Panagiotis Mitkidis.

**Project administration:** Janis H. Zickfeld.

**Validation:** Janis H. Zickfeld, Eyal Peer, Jareef Bin Martuza, and Yngwie A. Nielsen.

**Visualization:** Janis H. Zickfeld and Zoe Rahwan.

**Writing - original draft:** Janis H. Zickfeld.

**Writing - review & editing:** Janis H. Zickfeld, Karolina A. Ścigała, Christian T. Elbæk, John Michael, Mathilde H. Tønnesen, Gabriel Levy, Shahar Ayal, Eyal Peer, Eileen Chou, Ulrich Glogowsky, Simeon Schudy, Yuval Feldman, Patricia Kanngiesser, Nina Mazar, Simon Schindler, Alexa Weiss, Daniele Nosenzo, Valerio Capraro, Ralph Hertwig, Rachel Barkan, Dan Ariely, Ronit Montal-Rosenberg, Isabel Thielmann, Jareef Bin Martuza, Philipp Gerlach, Štěpán Bahník, Michał Białek, Jan K. Woike, Laila Nockur, Simen Bø, Nicolas Jacquemet, Nils Köbis, Sebastian Hafenbrädl, Zoe Rahwan, Agne Kajackaite, Yngwie A. Nielsen, Ann-Kathrin Koessler, Marek Vranka, Anna Z. Czarna, Alicia Seidl, Stefan Pfattheicher, and Panagiotis Mitkidis.

**Abstract**

Dishonest behaviors such as tax evasion impose significant societal costs. Ex-ante honesty oaths—commitments to honesty before action—have been proposed as useful interventions to counteract dishonest behavior, but the heterogeneity in findings across operationalizations calls their effectiveness into question. We tested 21 honesty oaths (including a baseline oath)—proposed, evaluated, and selected by 44 expert researchers—and a no-oath condition in a megastudy in which 21,506 UK and US participants played an incentivized tax evasion game. Of the 21 interventions, 10 significantly improved tax compliance by 4.5 to 8.5 percentage points, with the most successful nearly halving tax evasion. Limited evidence for moderators was found. Experts and laypeople failed to predict the most effective interventions, but experts' predictions were more accurate. In conclusion, honesty oaths can be effective in curbing dishonesty but their effectiveness varies depending on content. These findings can help design impactful interventions to curb dishonesty.

*Keywords:* honesty oath, dishonesty, tax compliance, nudging, unethical behavior

Words 150/150

# I Solemnly Swear I'm Up To Good: A Megastudy Investigating the Effectiveness of Honesty Oaths on Curbing Dishonesty

In ancient Greece, the oath was an institution crucial to everyday life, safeguarding social harmony and truth.[1] Swearing a false oath would incur punishment from the gods. In modern societies, oaths are still prevalent—for example, in courts of law (sworn testimonies), medicine (the Hippocratic oath), business administration (the MBA oath), and finance (the Dutch bankers' oath), with the main function of committing the oath-taker to a specific understanding of what is right or wrong in their professional context.[2,3] The prevalence and societal costs of dishonesty,[4] here defined as *distorting the facts as one sees them in order to acquire advantages or profits,* have led researchers to investigate the reasons for engaging in such behavior.[5–9] The literature suggests that ex-ante honesty oaths—committing to acting honestly before facing the temptation to transgress—may be effective in curtailing dishonesty.[10–12]

Psychological theories of dishonesty hold that people act dishonestly if they can justify doing so while maintaining a positive self-concept.[12,13] According to self-concept maintenance theory, honesty oaths should increase the salience of a moral standard, making it harder to justify dishonest acts while maintaining a positive view of the self.[12] In line with this reasoning, prior research suggests that honesty oaths may work by making individuals feel committed to telling the truth,[11,14] thus reducing self-justification processes.[12,15]

However, experimental studies have shown that some types of honesty oaths are ineffective[16–19] or even counterproductive.[20] Similarly, several field studies testing the effectiveness of honesty oaths or pledges to combat tax evasion, exam cheating, and insurance fraud have reported mixed or null findings.[21–23] In addition, recent cases of fraudulent research practices[24] and failed replications[17] have cast doubt on the effectiveness of honesty oaths.

Although a recent meta-analysis summarizing 124 effects[11] found that committing to honesty oaths, pledges, or honor codes can increase honesty (Cohen's $d = 0.27$, 95% CI = [0.19, 0.36]), it also revealed a high level of heterogeneity among the effects. In addition, the meta-analysis documented moderate evidence of publication bias, suggesting that the actual effects of such interventions may be smaller (smallest corrected effect: $d = 0.14$). At the same time, how people committed to oaths (e.g., signing, checking a box, verbalizing[25]), how the oath was operationalized, and the specific circumstances under which it was presented (e.g., directly before the critical response) varied considerably across studies, making it difficult to explain conclusively why some studies found positive effects and others did not. The question therefore remains: Under what conditions do honesty oaths make people behave more honestly, and to what extent do the mixed findings in the literature stem from variations in contexts or honesty oath operationalizations? The available evidence cannot separate variations in context and operationalizations, as they are rarely tested systematically. Given the substantial costs associated with dishonesty[26] and the high level of heterogeneity in the effects of honesty oath interventions, coupled with variation in how honesty oaths are operationalized, it is important to systematically and comprehensively evaluate the effectiveness of different types of honesty oath interventions.

Our study had five main objectives (Figure 1): We tested operationalizations of honesty oaths (i.e., oath formulation), the effects of commitment type (e.g., checking a box or retyping the oath), the effects of placement (e.g., directly before the target behavior or earlier), combinatorial effects of different psychological mechanisms (e.g., social norms, moral reminders), and individual, situational, and cultural moderators. To do so, we first crowdsourced possible honesty oath interventions from all collaborators, resulting in 98 suggestions. After several rounds of screening and voting, we narrowed the list to the 20 interventions collaborators expected to be the most effective and feasible to implement (for

details, see Supplementary Note 3). For comparison purposes we also added a control condition with no honesty oath and an intervention with a baseline honesty oath ("I hereby declare that I will provide honest information in this study"; see Table 1 for an overview of all interventions). We used a *megastudy* approach[27–29]—a large experimental study testing numerous interventions in the same population with the same outcome variable—to test the impact of the honesty oath interventions. In our megastudy, 21,506 participants in the UK and US played an incentivized game that allowed for dishonesty. We also asked collaborators, external behavioral scientists, finance experts, laypeople, and large language models (LLMs) to predict the outcome and the effectiveness of the honesty oath interventions in a forecasting survey.

Figure 1. Overview of the five main objectives of the study and their motivation. (See Supplementary Note 22 for detailed information on each objective.). Icons by flaticons.com, Adobe Stock, and GPT 4.0 Infographic Genius Pro.

We measured dishonesty using a tax evasion game (Figure 2). Tax evasion games have frequently been used in economic studies,[30] including in combination with honesty oaths,[14] and have demonstrated external validity.[31] These games model real-life tax reporting: Participants earn an income in a task (*actual income*) and self-report their earnings (*reported income*), which are then "taxed" (at 35% in the current task) to contribute to a collective resource. The ratio between reported and actual income specifies *tax compliance*. A compliance rate of 100% indicates an honest participant who fully pays due taxes, whereas a compliance rate of 0% indicates a fully dishonest participant. Participants in the honesty oath interventions were asked to commit to an honesty oath either before they began generating income or directly prior to reporting their income.

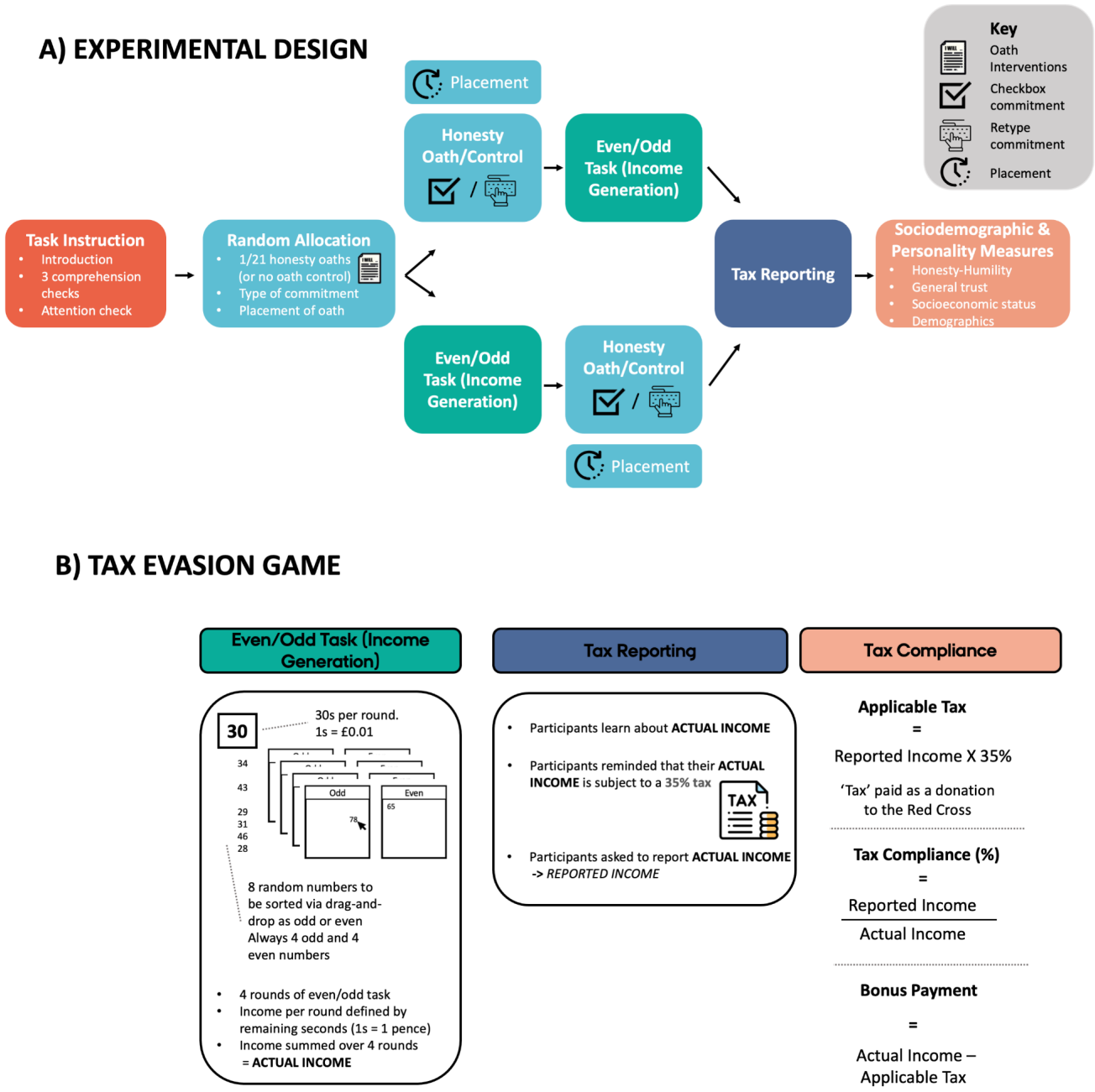


Figure 2. Overview of **(A)** experimental design and **(B)** tax evasion game employed in the main study. Icons by flaticons.com.

Table 1. Overview of the final interventions included in the main study

| # | Description | Formulation | *n* | Tax compliance rate (%: *M*, *SD*) | | Tax loss (£, % of total taxes) | | Dishonesty rate (%) | Fully Dishonest (%) | Tax compliance rate (%: *M*, *SD*) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Type of commitment | | Placement | |
| | | | Overall | | | | | | | Checkbox | Retype | Earlier | Directly Prior |
| -1 | Control condition | - | 953 | 82.3 | 33.5 | 41.5 | 21.9 | 31.3 | 9.4 | 83.5 (31.8) | 81.1 (35.1) | 82.4 (33.8) | 82.2 (33.2) |
| 0 | Baseline oath | **I hereby declare that I will provide honest information in this study.** | 955 | 85.0 | 31.0 | 35.3 | 18.1 | 27.5 | 8.4 | 84.8 (31.5) | 85.1 (30.7) | 82.9 (32.7) | 87.3 (29) |
| *Baseline reformulations (tax compliance: M = 87.2, SD = 28.8)* | | | | | | | | | | | | | |
| 1 | Specific behavior | **I hereby declare that I will provide honest information when reporting my final income from the sorting task.** | 999 | 89.7 | 26.3 | 25.3 | 11.6 | 18.5 | 4.8 | 86.9 (29.5) | 92.4 (22.6) | 88.6 (27.7) | 90.8 (24.8) |
| 2 | Severity of oath/honor | **I hereby swear upon my honor that I will provide honest information in this study.** | 939 | 84.5 | 31.1 | 35.1 | 18.5 | 28.1 | 7.2 | 83.6 (31.7) | 85.6 (30.4) | 81.1 (34.7) | 87.9 (26.6) |
| *Other-consequences: Harm (tax compliance: M = 86.2, SD = 29.9)* | | | | | | | | | | | | | |
| 3 | Harm/loss frame | **I understand that reporting dishonestly will decrease the amount of taxes and therefore money that goes to the Red Cross.** I hereby declare that I will provide honest information in this study. | 993 | 88.5 | 27.5 | 27.8 | 13.0 | 22.2 | 5.3 | 88.5 (27.3) | 88.6 (27.7) | 87.2 (29.1) | 89.8 (25.9) |
| 4 | Harm/gain frame | **I understand that reporting honestly will increase the amount of taxes and therefore money that goes to the Red Cross.** I hereby declare that I will provide honest information in this study. | 994 | 85.6 | 30.3 | 34.5 | 16.7 | 25.8 | 6.6 | 83.6 (32.5) | 87.5 (27.8) | 85.4 (30.3) | 85.7 (30.3) |
| 5 | Harm/gain frame/loyalty | **I understand that reporting honestly will increase the amount of taxes and therefore money that goes to the Red Cross, who are relying on loyal members of the community to support them.** I hereby declare that I will provide honest information in this study. | 980 | 85.7 | 30.0 | 33.5 | 16.3 | 26.1 | 6.3 | 86 (30) | 85.5 (30) | 84.5 (31.3) | 86.9 (28.6) |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | Harm/gain frame/people in need | **I understand that reporting honestly will increase the amount of taxes and help people in need and those worse off.** I hereby declare that I will provide honest information in this study. | 1003 | 87.7 | 28.0 | 30.7 | 14.3 | 22.5 | 5.1 | 87.7 (28.6) | 87.8 (27.4) | 88.1 (27.4) | 87.4 (28.7) |
| 7 | Societal loss | **In general, tax avoidance results in less funding for schools, hospitals and welfare. It hurts me and the society.** I hereby declare that I will provide honest information in this study. | 956 | 84.9 | 31.4 | 35.4 | 18.2 | 26.6 | 8.7 | 85.4 (31.6) | 84.5 (31.2) | 85 (30.8) | 84.8 (32) |
| 8 | Collective evasion | **I understand that even small misreports by single participants will add up across participants.** I hereby declare that I will provide honest information in this study. | 1032 | 84.6 | 31.9 | 40.0 | 19.0 | 26.9 | 8.8 | 84.1 (32.1) | 85 (31.7) | 81.7 (34.4) | 87.5 (28.7) |
| *Other-consequences: Social bonds (tax compliance: M = 84.1, SD = 31.9)* | | | | | | | | | | | | | |
| 9 | Sense of community | **We are in this together. Accurate tax reporting is important for a functioning society.** I hereby declare that I will provide honest information in this study. | 944 | 85.5 | 31.0 | 34.1 | 17.4 | 24.4 | 7.4 | 84.4 (32.1) | 86.5 (29.8) | 85.4 (30.3) | 85.5 (31.7) |
| 10 | Trust | **To earn the trust of my fellow citizens,** I hereby declare that I will provide honest information in this study. | 1027 | 82.7 | 33.5 | 44.4 | 21.2 | 28.8 | 9.5 | 82.7 (32.8) | 82.8 (34.2) | 81.3 (34.5) | 84.1 (32.4) |
| 11 | Empathy | **By being honest, I show compassion and empathy for others and society.** I hereby declare that I will provide honest information in this study. | 980 | 84.2 | 31.2 | 37.7 | 18.9 | 28.4 | 7.4 | 83.5 (32.3) | 84.9 (29.9) | 82.3 (32.1) | 86.1 (30.1) |
| *Description of dishonesty/situational (tax compliance: M = 88.0, SD = 27.8)* | | | | | | | | | | | | | |
| 12 | Meaning of dishonesty | **I understand that dishonest reporting is a fraudulent way of getting money I do not deserve.** I hereby declare that I will provide honest information in this study. | 1027 | 88.2 | 27.9 | 29.7 | 13.7 | 22.2 | 6.2 | 88 (27.9) | 88.3 (27.9) | 86.5 (29.3) | 89.7 (26.4) |
| 13 | Binary dishonesty | **I understand that honesty is an all-or-nothing concept: Either the reporting is honest or it is not.** I hereby declare that I will provide honest information in this study. | 984 | 86.6 | 29.6 | 32.7 | 16.1 | 24.6 | 7.1 | 86.3 (30.2) | 86.8 (28.9) | 84.8 (31.1) | 88.2 (28.1) |
| 14 | Misreporting forbidden | **I understand that misreporting is forbidden in this study.** I hereby declare I will provide honest information in this study. | 993 | 89.4 | 25.9 | 25.9 | 12.0 | 21.5 | 4.8 | 88.3 (27) | 90.4 (24.7) | 88.7 (26.7) | 90 (25.1) |
| *Self-consequences: Self-image (tax compliance: M = 85.5, SD = 30.7)* | | | | | | | | | | | | | |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 15 | Responsibility | **I understand that it is my responsibility to report honestly.** I hereby declare that I will provide honest information in this study. | 918 | 86.2 | 30.4 | 30.2 | 15.8 | 24.4 | 7.5 | 83.5 (32.5) | 88.8 (28) | 85.8 (30.6) | 86.6 (30.1) |
| 16 | Character | **I am an honest person and** therefore hereby declare that I will provide honest information in this study. | 975 | 84.7 | 31.1 | 35.8 | 17.9 | 27.8 | 7.7 | 84.3 (31.3) | 85 (30.9) | 84.2 (31.4) | 85.1 (30.8) |
| 17 | Guilt avoidance | **When reporting honestly I will avoid feeling guilty afterwards.** I hereby declare that I will provide honest information in this study. | 963 | 85.8 | 30.6 | 34.5 | 17.2 | 25.0 | 6.9 | 86.4 (30.8) | 85.1 (30.3) | 84.7 (32) | 87.0 (28.9) |
| *Social norms (tax compliance: M = 86.8, SD = 29.4)* | | | | | | | | | | | | | |
| 18 | Injunctive norm | **I understand that most people agree that reporting honestly is the right thing to do.** I hereby declare that I will provide honest information in this study. | 1013 | 87.0 | 29.0 | 32.8 | 15.3 | 24.3 | 6.4 | 88.5 (27.3) | 85.4 (30.5) | 85.1 (31.2) | 88.9 (26.5) |
| 19 | Descriptive norm | **I understand that based on previous similar studies, 89% of participants reported honestly.** I hereby declare that I will provide honest information in this study. | 957 | 86.7 | 29.8 | 32.0 | 15.8 | 23.4 | 7.0 | 87.4 (29.6) | 86 (30) | 86.3 (29.4) | 87.1 (30.2) |
| *AI-generated comparison (tax compliance: M = 87.0, SD = 29.9)* | | | | | | | | | | | | | |
| 20 | Chat GPT | **I solemnly pledge to report my income truthfully, directly benefiting the British Red Cross and those in need. I recognise my honesty fosters a culture of trust, mutual respect, and social responsibility, inspiring others to act with integrity. By upholding these values, I contribute to building a just and equitable society for all.** | 921 | 87.0 | 29.9 | 29.1 | 15.1 | 22.3 | 6.7 | 86.4 (30.3) | 87.6 (29.5) | 85.7 (31.7) | 88.3 (28.1) |

"Earlier" indicates that the honesty oath was completed in the beginning before the income generation task. "Directly prior" indicates it was completed after the income generation task and directly before reporting income. Categorisations of honesty oath interventions detailed in Supplementary Note 15.

## Results

An overview of all registered hypotheses, exploratory questions, analyses, and main findings is presented in Table 2. For all analyses using equivalence testing (i.e., to test for the absence of an effect) we set our smallest effect size of interest at $d = 0.15$ (~ 5 percentage points; see Supplementary Note 14). For all analyses, we set the alpha level at 0.05. All confidence intervals refer to 95% confidence intervals unless stated otherwise.

Table 2. Overview of confirmatory hypotheses and exploratory questions, main analyses, and findings

| # | **Hypothesis/Question** | **Goal** | **Analyses** | **Finding** |
|---|---|---|---|---|
| | **Confirmatory** | | | |
| H1 | Commitment to an honesty oath increases tax compliance compared to the control condition including no honesty oath. | Comparing honesty oath (all oaths together) to the control (i.e., no oath) | DV: TC;<br>IV: honesty oath (no oath −0.5, oath 0.5)[a] | **Confirmed.** *OR* = 1.18 [1.06, 1.30], *d* = 0.09 |
| | | Comparing all individual honesty oaths to the control | DV: TC<br>IV: Type of honesty oath (comparing all interventions to control)[a] | **Confirmed.** 10/21 statistically significant (when controlling for multiple comparisons) |
| | | Adjusting for potential winner's curse | James-Stein shrinkage | Most effective intervention: *OR* = 1.36, *d* = 0.17 |
| H2 | Tax compliance differs across the different honesty oaths. | | Multiple comparisons | **Confirmed.** $\chi^2(21) = 69.56$, $p < .001$ |
| H3 | There is no difference between committing via checking a box or retyping the statement on misreporting in the tax evasion game. | | DV: TC<br>IV: Type of commitment (checkbox: −0.5, retype: 0.5)[a] | **Confirmed.** No statistically significant main effect (*OR* = 1.04 [1.00, 1.09], *d* = 0.02); smaller than smallest effect size of interest. |
| | | Type of commitment as a moderator | DV: TC<br>IV: Type of honesty oath × type of commitment [a] | No significant interaction effect, $\chi^2(21) = 24.36$, $p = .276$ |
| H4 | There is no difference in completing the | | DV: TC | **Confirmed.** Statistically significant main effect (*OR* = 1.13 |

| | | | | |
|---|---|---|---|---|
| | oath, before or after the even/odd task. | | IV: Placement (before: −0.5, after: 0.5)[a] | [1.08, 1.19], $d = 0.07$); but smaller than smallest effect size of interest |
| | | Placement as a moderator | DV: TC<br>IV: Type of honesty oath × placement [a] | No significant interaction effect, $\chi^2(21) = 14.13$, $p = .864$ |
| | **Exploratory** | | | |
| | Do intervention effects differ by trait Honesty-Humility? | | DV: TC<br>IV: Type of oath × Honesty-Humility (centered)[a] | Significant main effect ($\tau = 0.14$, $p < .001$); no significant interaction, $\chi^2(21) = 27.87$, $p = .144$. Significant interaction for #9 |
| | Do intervention effects differ by trait general trust? | | DV: TC<br>IV: Type of oath × trust [a] | Significant main effect ($\tau = 0.07$, $p < .001$), no significant interaction, $\chi^2(21) = 29.58$, $p = .101$. Significant interaction for #9 |
| | Do intervention effects differ by socioeconomic status? | | DV: TC<br>IV: Type of oath × socioeconomic status [a] | No significant main effect, no significant interaction, $\chi^2(21) = 28.57$, $p = .125$. Significant interaction for #3, #18. Significant main effect without interaction ($\tau = -0.02$, $p < .001$) |
| | Do intervention effects differ by location? | | DV: TC<br>IV: Type of oath × location [a] | Significant main effect (U.S.: $M = 82.8\%$, $SD = 33.0\%$, $n = 7{,}705$; U.K.: $M = 87.80\%$, $SD = 28.2\%$, $n = 13{,}801$; $OR = 1.28$ [1.22, 1.34], $d = 0.14$, $p < .001$); no significant interaction, $\chi^2(21) = 16.54$, $p = .738$ |
| | Do intervention effects differ by gender? | | DV: TC<br>IV: Type of oath × gender (including female and male) | Significant main effect (female: $M = 88.4\%$, $SD = 26.9\%$, $n = 12{,}069$; male: $M = 83.1\%$, $SD = 33.4\%$, $n = 9{,}047$; $OR = 1.37$ [1.12, 1.68], $d = 0.17$, $p = .002$); significant interaction, $\chi^2(21) = 34.38$, $p = .034$ |

| | | | | |
|---|---|---|---|---|
| | Do intervention effects differ by age? | | DV: TC<br>IV: Type of oath × age | No significant main effect; no significant interaction, $\chi^2(21) = 26.62$, $p = .184$. Significant main effect without interaction ($\tau = 0.05$, $p < .001$) |
| | Are interventions more effective than the baseline oath? | | DV: TC<br>IV: Type of honesty oath (comparing all interventions to base oath)[a] | 0/20 statistically significant (when controlling for multiple comparisons) |
| | Are different intervention types more effective than the control (i.e., no oath)? | | DV: TC<br>IV: Type of intervention (vs. control)[a] | 6/8 statistically significant; description/situational most effective, social bond least effective |
| | Are different intervention types as grouped by the DENIAL framework more effective than control? | | DV: TC<br>IV: DENIAL [a] | 6/8 statistically significant; intervention manipulating all aspects most effective, social bond least effective |
| | Is time spent on tasks associated with tax compliance? | Time spent on oath | DV: TC<br>IV: Time spent on oath [a] | Significant effect; $\tau = 0.03$, $p < .001$ |
| | | Time spent on income reporting | DV: TC<br>IV: Time spent on income reporting [a] | Significant effect: $\tau = -0.12$, $p < .001$ |
| | Is the effect of honesty oaths on tax compliance mediated by time spent on income reporting? | | DV: TC<br>IV: Honesty oath<br>M: Time spent on income reporting [a] | Significant indirect effect: 0.03 [0.02, 0.04] |
| | **Forecasting survey** | | | |
| | How well can forecasters predict relative effectiveness? | | Observed/average marginal effects with predicted effects per population[b] | Small correlations; collaborators show the worst relative predictions |
| | How well can forecasters predict the actual effects of interventions? | | DV: Absolute deviation<br>IV: Population[c] | Significant differences; collaborators, behavioral scientists and LLMs show smallest error; significantly smaller than general population and finance experts |

*Note*: DV = dependent variable. IV = independent variable. M = mediator. TC = tax compliance. Analysis models: [a] ordered beta regression; [b] correlational analysis; [c] multilevel regression.

**Descriptive Results**

Overall, we observed an average tax compliance of 86.0% (*SD* = 30.1%; Supplementary Figure 16). A total of 5,398 (25.1%) participants underreported their income to some degree, of which 1,519 participants falsely reported no income (7.1% of total sample; 28.1% of dishonest participants) in order to evade all possible taxes. A total of £737.96 was lost due to tax evasion (14.2% of all taxes). We observed only a small negative correlation between actual income and tax compliance ($r = -0.02$ 95%, CI [−0.04, −0.01]), which did not provide credible evidence when testing it against the smallest effect size of interest (Supplementary Note 14).

**Confirmatory Results**

**H1. Commitment to an honesty oath increases tax compliance.** Across all interventions, including an oath increased average tax compliance by 3.9 percentage points (*M* = 86.2%, *SD* = 29.9%, *n* = 20,553) compared to the control condition of no oath (*M* = 82.3%, *SD* = 33.5%, *n* = 953), *OR* = 1.18 [1.06, 1.30], *d* = 0.09, *p* = .002. All oath interventions showed higher mean tax compliance than the control condition (Table 1; Figure 3). To account for multiple comparisons, we adjusted all *p*-values using the Benjamini-Hochberg procedure.[28] Considering our alpha level of .05, we found that 10 of 21 oath interventions showed higher tax compliance compared to the control, increasing tax compliance by 4.5 to 8.5 percentage points (see Figure 3, Supplementary Table 15).

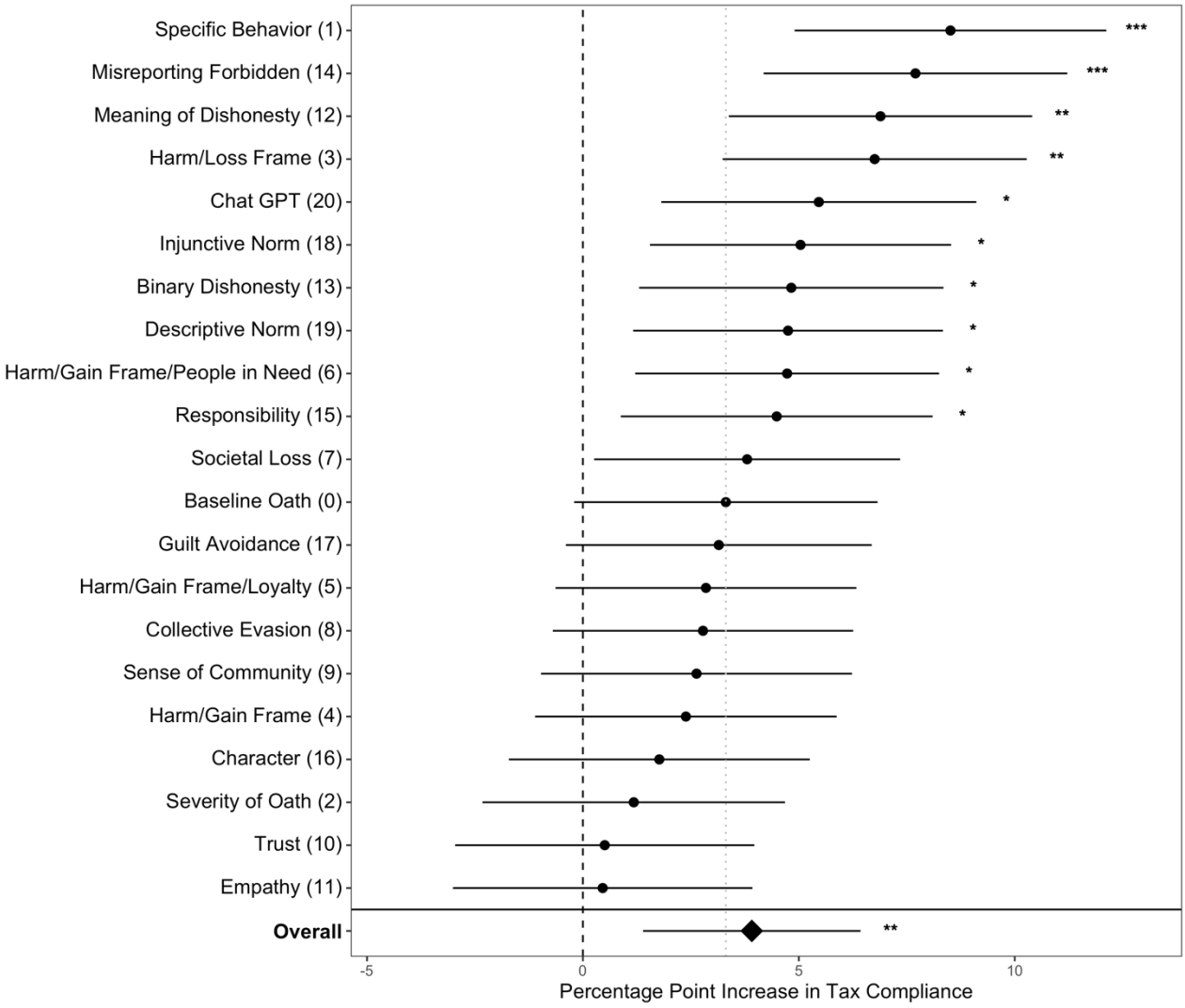


Figure 3. Overview of average marginal effects for oath interventions compared to the control (no oath) condition (black dashed vertical line) in percentage point increases in tax compliance. Baseline oath level is shown as a gray dotted vertical line for comparison. Whiskers depict 95% CIs without correction for multiple comparisons. Asterisks define statistically significant interventions (based on adjusted *p*-values): * $p_{adj} < .05$, ** $p_{adj} < .01$, *** $p_{adj} < .001$.

The most effective intervention, which referred to the specific behavior in the task ("I hereby declare that I will provide honest information when reporting my final income from the sorting task"; #1 Specific Behavior), showed an increase of 8.5 percentage points (*OR* = 1.43 [1.23, 1.67]; $d = 0.20$, $p < .001$). In the control intervention, more than a fifth of due taxes were lost (21.9%); the most effective intervention reduced this loss to 11.6%, thus nearly cutting tax losses in half (47.0%; Table 1). Similarly, the control intervention showed a dishonesty rate of 31.3%, suggesting that nearly a third of participants misreported their

income, and 9.4% of participants reported no income at all in order to evade all potential taxes. The most effective intervention nearly halved both the dishonesty rate (18.5%) and the number of participants falsely reporting no income (4.8%; Table 1).

To control for a possible overestimation of effects due to the *winner's curse*—the likelihood that the largest effects are overestimated—we applied the James-Stein shrinkage procedure.[29] The strength of the most effective intervention, Specific Behavior, was only slightly reduced (from $OR = 1.43$, $d = 0.20$, 8.5 percentage points to adjusted $OR = 1.36$, $d = 0.17$, 7.4 percentage points), similar to the other interventions (Supplementary Table 17). The findings were also robust when controlling for participants' location (US vs. UK), type of commitment to the oath (checkbox vs. retyping), oath placement (earlier vs. directly prior tax reporting), device type (mobile vs. desktop), gender, and age (Supplementary Table 20).

We performed equivalence tests to investigate whether intervention effects were equivalent to zero and observed that four interventions (#2 Severity of Oath; #10 Trust; #11 Empathy; #16 Character; see Table 1)—did not show credible evidence compared to the control (Supplementary Table 16).

**H2. Tax compliance differs across honesty oath interventions.** We observed a difference across honesty oath interventions in tax compliance, $\chi^2(21) = 69.56$, $p < .001$, suggesting that honesty oath interventions differed in their effectiveness (Supplementary Table 21).

**H3. There is no substantial difference in misreporting between checking a box and retyping the oath.** Overall, tax compliance tended to be higher for participants who retyped the honesty oath ($M = 86.7\%$, $SD = 29.4\%$, $n = 10{,}319$) than for participants who checked a box ($M = 85.8\%$, $SD = 30.5\%$, $n = 10{,}234$), but the difference was not statistically significant ($OR = 1.04$ [1.00, 1.09], $d = 0.02$, $p = .065$) and was significantly smaller than our smallest effect size of interest (Supplementary Note 14). However, two honesty oath

interventions, Specific Behavior (increase in percentage points compliance rate 5.0 vs. 12.6) and Responsibility ("I understand that it is my responsibility to report honestly."; 0.6 vs. 8.6 percentage points), were significantly more effective when participants retyped them rather than when they checked a box (Supplementary Tables 22, 23; Supplementary Figure 17).

**H4. There is no substantial difference in effectiveness based on placement of the honesty oath.** Placement had a significant effect on tax compliance, however, significantly smaller than our smallest effect size of interest (Supplementary Note 14). Tax compliance was higher if the oath was completed after the income-generating task and directly prior to tax reporting ($M$ = 87.4%, $SD$ = 28.8%, $n$ = 10,328) rather than earlier before the task ($M$ = 85.0%, $SD$ = 310%, $n$ = 10,225; $OR$ = 1.13 [1.08, 1.19], $d$ = 0.07, $p < .001$). Completing the oath directly prior to reporting on average increased tax compliance by 3.0 percentage points. In tests for whether this effect differed for the 21 honesty oaths, we found no statistically significant interaction effects ($\chi^2(21) = 14.13$, $p$ =.860; Supplementary Table 24). In general, effects were stronger when the honesty oath appeared directly prior to tax reporting rather than earlier (Supplementary Table 25; Supplementary Figure 17).

**Exploratory Results**

We performed several exploratory analyses focusing on possible moderation by personality (Honesty-Humility, general trust), economic (subjective socioeconomic status), demographic (gender, age), and cultural (location) variables (see Table 2, Supplementary Note 15). We observed positive correlations for Honesty-Humility, general trust, and participant age with tax compliance and a small negative correlation of subjective socioeconomic status with tax compliance (see Table 2, Supplementary Table 26). Participants in the US showed lower tax compliance on average compared to participants in the UK and male participants showed lower tax compliance on average compared to female participants (Table 2). Adding an honesty oath reduced the difference in tax compliance

between the categories for both location and gender (Supplementary Note 15). There were no significant moderation effects of the effects of honesty oath interventions on tax compliance by Honesty-Humility, general trust, subjective socioeconomic status, location, or participant age (Table 2). For these variables, only a small number of individual honesty oath interventions showed an interaction effect (see Supplementary Note 15). We found a statistically significant moderation effect by gender: Due to ceiling effects in tax compliance for female participants, honesty oaths were, on average, more effective for male participants (Supplementary Table 44).

We also tested whether oath interventions differed statistically significantly from the baseline oath intervention. We observed a statistically significant effect for two honesty oath interventions, Specific Behavior (increase of 5.2 percentage points over the baseline oath) and Misreporting Forbidden (increase of 4.4 percentage points over the baseline oath), but not when adjusting for multiple comparisons (Supplementary Tables 32, 33). Additional exploratory results and analyses on categorisations of honesty oath interventions and time spent is provided in Table 2 and Supplementary Note 15.

**Forecasting Survey**

We investigated how well five samples were able to predict the current findings in absolute percentage points: collaborators of the current project (scientists considered experts in behavioral science or economics; $n = 34$), external behavioral scientists (behavioral scientists not taking part in the current project; $n = 29$); laypeople ($n = 167$), experts working in the finance and insurance industry ($n = 99$), and LLMs (ChatGPT 4.0, Llama 2, Bard; $n = 3$).

To test the relative predictive power, we computed correlation coefficients between the observed effects for each intervention and the average predicted effects for each intervention per sample. We observed small, nonsignificant correlations. If anything,

collaborators ($r_{21}$ = 0.07 [−0.37, 0.49]) were worse at predicting the relative effectiveness compared to external behavioral scientists ($r_{21}$ = 0.17 [−0.28, 0.56]), laypeople ($r_{21}$ = 0.18 [−0.28, 0.56]), finance and insurance experts ($r_{21}$ = 0.18 [−0.28, 0.56]), and LLMs ($r_{21}$ = 0.16 [−0.29, 0.55]; for sensitivity with rank-order correlations and correlations with average marginal effects instead of observed effects, see Supplementary Table 47). We performed a sensitivity analysis to account for potential wisdom of crowd effects (i.e., more accurate predictions due to higher sample size), but found no differences due to sample size (Supplementary Note 19).

The predictions of both collaborators and external behavioral scientists showed the smallest mean absolute deviation from the observed effects (collaborators: $M$ = 3.54, $SD$ = 2.70; external behavioral scientists: $M$ = 3.99, $SD$ = 3.34) compared to laypeople ($M$ = 8.82, $SD$ = 7.35) and experts in finance and insurance ($M$ = 8.30, $SD$ = 7.34), but were not statistically significantly different from LLMs ($M$ = 4.22, $SD$ = 3.10; see Supplementary Figures 33, 35, Supplementary Table 48; of LLMs, ChatGPT 4.0 performed the worst; see Supplementary Figures 36, 37). These findings were replicated when focusing on the absolute deviation from the average marginal effects (Supplementary Table 50). We also observed that the mean absolute deviation differed among interventions (Supplementary Figure 38; Supplementary Tables 49, 51).

## Discussion

We investigated the effectiveness of ex-ante honesty oaths in reducing dishonesty by crowdsourcing interventions from experts in the field and testing them across 21,506 US- and UK-based participants online. Overall, commitment to an honesty oath reduced subsequent misreporting of taxes more than no oath (i.e., when honesty was not made salient in any other way; control). We observed considerable heterogeneity across interventions, with honesty oaths increasing tax compliance by 0.5 to 8.5 percentage points. After adjusting for multiple comparisons, of the 21 interventions (including the baseline oath), 10 significantly increased tax compliance from the control condition: highlighting honesty in income reporting, emphasizing that misreporting is forbidden, clarifying the meaning of dishonesty, highlighting the harm of dishonesty using a loss frame, an AI-generated statement (included to distinguish effects from human-generated statements) emphasizing several aspects, highlighting an injunctive norm, highlighting a descriptive norm, clarifying the binary nature of dishonesty ("you can either be honest or not"), highlighting the harm done to people in need, and appealing to individual responsibility. The five most effective interventions also increased tax compliance by 2.2 to 5.2 percentage points from the baseline oath. Yet, none of them was significantly different from the baseline oath when adjusting for multiple comparisons. Interventions were mostly robust when considering location of the study (UK vs. US), type of commitment (checkbox vs. retyping), placement (before vs. after income-generation task), and personality traits (Honesty-Humility, general trust).

### Effectiveness of Honesty Oaths

Honesty oaths that specified procedures or made situations and behaviors less ambiguous were among the most effective in comparison to control. Ambiguity and the potential to attribute a dishonest act to the environment are important predictors for individuals engaging in dishonest behavior.[15,32,92] Reducing a situation's ambiguity by

emphasizing the specific behavior targeted by the oath (e.g., reporting one's final income), making rules less ambiguous (e.g., stating that misreporting income is forbidden), and explaining and defining honesty (e.g., describing honesty as an all-or-nothing behavior) were all effective in reducing misreporting. Therefore, offering not only the opportunity to commit to telling the truth but also a definition of what truth means in that context might increase the effectiveness of honesty oaths. These findings align with theories in motivational psychology that highlight the effectiveness of reducing goal ambiguity for goal achievement,[33,34] as well as with findings that specific rules have a stronger impact on reducing dishonesty than general rules.[35,36] These insights also go beyond interventions focusing on moral reminders that have shown mixed effects.[37–39] We did not test whether honesty oaths were more effective than the same description without a commitment (i.e., a moral reminder). Previous studies found evidence that honesty oaths can be more effective than moral reminders,[40] but future research should test this systematically.

Honesty interventions focusing on social norms were also effective in reducing dishonesty. This replicates previous findings[41–46] suggesting that providing information about peer attitudes and behavior can mitigate dishonest reporting. In the current study, two social norms—an injunctive norm that focused on what most people would generally approve of and a descriptive norm that highlighted the actual behavior of previous participants—were effective. This finding is in contrast to previous experimental and field studies that have observed no effects of social norms on reducing dishonesty.[47–49,93] It is possible that committing to an honesty oath highlighting a social norm helps a person internalize the norm, making this type of intervention more effective than simply providing information on a social norm. Future studies should test the effectiveness of presenting social norms with and without commitment to honesty oaths.

An AI-generated honesty oath was effective in increasing tax compliance, speaking to a broader trend of using emerging technologies to curb unethical behavior.[50,51] It is difficult to disentangle the exact mechanism of this intervention, since the AI-generated oath manipulated several factors at the same time by reducing ambiguity around the targeted behavior, appealing to potential benefits of paying taxes, and highlighting social responsibility.

The least effective interventions focused on potential consequences for the participant or self-image effects and those highlighting a social bond. Appealing to the participant's moral character or highlighting that honesty is important for trust and a functioning society did little to increase tax compliance. This observation is counterintuitive for two reasons. First, theories of dishonesty highlight that self-image and social-image concerns are important in reducing dishonesty.[12,52] Second, misreporting in our study had direct negative consequences on funds allocated to a charity whose work was to foster societal well-being. One possible explanation is that in anonymous or nonsocial settings such as our experiment or when reporting income taxes, there are limited opportunities to impress or gain the trust of others. Interventions highlighting self-image or social-image concerns might therefore be more effective in public decisions.

Importantly, while the most effective interventions increased tax compliance compared to the baseline oath (i.e., an honesty declaration focusing on the most basic aspects of honesty oaths and not including additional variables) by up to 5.2 percentage points, none was statistically significant when accounting for multiple comparisons. This finding suggests a potential power issue when contrasting smaller effects and we need to acknowledge that future well-powered studies would need to test whether the formulation of honesty oaths can have additional effects over a baseline formulation.

**Moderators of Honesty Oaths' Effectiveness**

In general, the relative effectiveness of honesty oath interventions was robust across contexts and personality variables. Placement, location, gender, and personality traits impacted tax compliance, but significantly moderated only some of the honesty oath interventions.

While we replicated findings from a recent meta-analysis suggesting that how people commit to the oath plays a less important role,[11] retyping the oath was more effective than checking a box for some of the interventions, replicating previous results.[25] The most effective interventions were more effective when participants retyped the oath than when they checked a box.

Honesty oath interventions presented directly before the opportunity allowing for misreporting were more effective than interventions presented at the beginning of the task, well before reporting, but this effect was not substantial based on comparison to our smallest effect size of interest. Overall, this effect was similar across interventions and suggests that connecting the honesty oath as closely as possible to the behavior to be influenced might be helpful. This observation also calls into question how long the effects of honesty oaths last; future studies are needed to evaluate this issue (for one relevant study see [25]).

Tax compliance differed between the UK and the US, with UK-based participants showing higher compliance on average. The honesty oath interventions had stronger effects in the US-based sample, possibly because the baseline compliance level was lower. Tax systems, attitudes towards taxes, and tax gaps differ across countries and cultures, which might influence the general motivation to misreport taxes.[53] Nevertheless, we did not observe a statistically significant moderation by country, suggesting that the overall effect of honesty oaths could replicate in different tax contexts. Research on the cross-cultural differences of honesty oaths are limited, with only a few of the reviewed articles in a recent meta-analysis

directly comparing two countries or more.[11] Systematic research on the effectiveness of honesty oaths across cultural contexts is needed.

We observed some indication of moderations by Honesty-Humility, general trust, and gender for individual oaths. Whereas overall tax compliance increased with higher levels of Honesty-Humility and general trust,[54] some interventions were more effective for participants reporting low levels of these variables—most likely because individuals high in Honesty-Humility and general trust were less likely to misreport in the first place, thus reducing the potential effectiveness of any type of honesty intervention. Similarly, whereas female participants showed higher tax compliance on average replicating previous meta-studies[6,94], many honesty oath interventions were more effective for male participants, possibly driven by the fact that male participants misreported more in the control intervention than did female participants. Taken together, these findings highlight the potential of honesty oaths to curb dishonesty particularly in contexts or in populations that bring the risk of elevated dishonesty.

**Forecasting Results**

None of the forecasting samples correctly predicted the best-performing intervention. Similar to previous megastudies,[28,29] we observed low predictive power of relative effects across all samples. This underlines the value of investigating many interventions at once, since it may be difficult to predict beforehand which interventions will ultimately work best. Behavioral scientists, laypeople, and individuals working in the finance and insurance industry were slightly better at predicting the relative order of effects than collaborators, whereas collaborators and external behavioral scientists were more accurate at predicting the size of the actual effects than were laypeople and individuals working in finance and insurance. These findings are similar to those in studies comparing predictions of experimental results by experts and laypeople.[55] Artificial intelligence was broadly in line

with the behavioral scientists (collaborators and external) in terms of predictions, but showed low reliability across LLMs.

**Limitations and Constraints on Generalizability**

Although we provided a large-scale investigation of the effectiveness of honesty oaths on dishonesty, there are limitations to the current study. First, we employed a bottom-up approach to generate honesty oath interventions. Experts likely based their suggestions on theoretical and empirical knowledge in the field, but we did not employ a unified theoretical framework to generate interventions. Our agnostic approach is common for megastudies,[27] and we attempted to go beyond comparing interventions by exploring them according to theoretical frameworks. Nevertheless, we could not systematically test specific mechanisms to assess why some honesty oaths were more effective than others; future studies are needed to investigate whether commitment or cognitive dissonance reduction might be important for the effectiveness of honesty oaths.

Second, compared to previous studies testing honesty oaths, we observed small standardized effects. A recent meta-analysis observed an average effect of *Cohen's d* = 0.27 [0.19, 0.36], whereas our study found an overall effectiveness of $d = 0.09$ ($d = 0.20$ for the most effective intervention).[11] Based on a simulation study, we expected to be able to detect effects as small as $d = 0.11$ with 95% power. Therefore, some caution is warranted when interpreting smaller effects, as the current design might not be sufficiently powered to detect them. A recent series of studies also suggested that the effect of honesty oaths on dishonesty was small ($d = 0.13$)[56] and publication bias corrections reduced the findings of a recent meta-analysis to $d = 0.14$.[11] This highlights the importance of what has been traditionally referred to as *small* effect sizes[57] and suggests that future studies testing honesty oaths need to be adequately powered to detect effect sizes at such magnitude. Nevertheless, whereas standardized effects were small in comparison to previous studies, unstandardized effects

were of practical significance, with interventions increasing tax compliance up to 8.5 percentage points. The most effective intervention was able to cut tax losses almost in half, which is of high practical relevance given the cost effectiveness and wide applicability of implementing honesty oaths.

Third, our study's applicability and generalizability might be specifically tied to the characteristics of the targeted populations and the context of the tax evasion game. Although we tested two countries, it is unclear how generalizable the effects are to other countries or cultural contexts. Would honesty oaths focusing on the environment and reducing ambiguity be the most effective across different contexts? Would cultural factors such as a focus on collectivism influence the effectiveness of honesty oaths that emphasize social bonds? Our study also does not provide information on its generalizability beyond the current game context. Although we tried to model a real-life tax reporting setting as closely as possible by employing externalities that mirror societal outcomes of taxes, the setting has limited ecological validity. It is unclear whether the employed honesty oaths would produce similar effects in more applied contexts, such as reporting income taxes. Importantly, taxes and incomes are higher in real life contexts compared to the current experimental settings. Nevertheless, our insights can be important for designing interventions in such applied settings. Similarly, previous studies have shown that honesty oaths can be effective across a range of outcome behaviors including dishonesty in economic game tasks,[58–61] preference elicitation,[62,63] tax compliance,[22] and cheating in online exams.[64] Our study replicates this general finding and provides an overview of boundary conditions and effective formulations.

**Practical Recommendations for Ex-Ante Honesty Oath Interventions**

Despite the study's limitations, it provides evidence that can inform preliminary practical recommendations for designing effective honesty oath interventions (Figure 4). Ideally, each of these recommendations would be systematically evaluated and replicated in

further studies; many aspects, however, are already supported by previous empirical findings.[6] Recommendations include focusing explicitly on the target behavior (e.g., referring to the action that might be subject to dishonesty), defining the meaning of honesty, requiring relatively high involvement (e.g., retyping or signing), setting the honesty oath in close proximity to the target behavior, and targeting specific populations (e.g., younger males low in Honesty-Humility).


Figure 4. Practical recommendations for designing effective ex-ante honesty oath interventions. Recommendations are based on the current and previous meta-analyses[6,7,11]

and need to be evaluated systematically in future studies to ascertain their effectiveness.

Icons by flaticons.com

## Conclusion

We crowdsourced honesty oath interventions and tested them across a large-scale sample of online participants, finding evidence that honesty oaths can be effective in reducing dishonesty—in this case, misreporting in a tax evasion game. The most effective interventions, which highlighted the specific behavior and situation, emphasized social norms, and underscored potential costs to others, reduced dishonesty rates and tax losses by as much as nearly 50%. Importantly, we found that the effectiveness of honesty oaths depends on their specific formulation. While other honesty interventions, such as audits or punishment, might be effective in specific contexts, the current study offers evidence that honesty oaths can serve as low-key, cost-effective interventions to curb dishonesty.

## Methods

The current project was divided into three steps: First, we contacted collaborators and crowdsourced possible interventions. We also conducted three pilot studies to test the validity of the main task (Supplementary Note 8,9, and 10). Second, we ran the main study based on a final selection of interventions. Third, we ran a forecasting survey asking collaborators, laypeople, experts, and LLMs to predict the outcomes of the main study. All studies were reviewed and approved by the ethical review board of Aarhus University (BSS-2023-018; BSS-2023-106). We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Protocols and analysis plans for two of the pilot studies, the intervention selection, the main study, and the forecasting study were preregistered at https://osf.io/t3sm4/registrations. All deviations from these protocols are

provided in detail in Supplementary Notes 5, 13, and 17. All datasets and syntax are available at https://osf.io/t3sm4/.

**Crowdsourcing Interventions**

The project followed a consortium approach,[65] inviting experts to contribute to the project by suggesting possible interventions and providing funding. The first part of the project focused on recruiting experts as collaborators and having them suggest and vote on interventions.

**Participants.** The first and last authors assembled a core group of nine researchers. After an initial general meeting, the core group contacted potential collaborators based on their expertise in the field. Of the 52 researchers contacted, 35 confirmed that they would take part in the project. Of the 44 overall, three researchers left during the course of the project, for a final project group of 41 collaborators.

**Procedure.** After being invited to the project, collaborators were able to comment on the project description (see Supplementary Note 1). As a first step, collaborators were asked to suggest possible interventions. There was no limit on the number of interventions, but it was recommended that collaborators don't submit more than three interventions to keep the final number manageable. Collaborators were provided with the project description and detailed information on the type of interventions they could suggest, as well as with information on the potential design of the study (e.g., that the main study includes a control intervention without an honesty oath and a base honesty oath intervention that read, "I hereby declare that I will provide honest information in this study"; see Supplementary Note 2 for how this base formulation was derived). Collaborators were tasked with suggesting additional interventions that change the base formulation of the honesty oath (e.g., by using "I hereby swear" instead of "I hereby declare"), add information to the base formulation (e.g., stating the negative consequences of dishonesty), or change aspects of the design or outcome variable (e.g., adding the possibility of punishment). Intervention suggestions were not

anonymous so that potential questions about the interventions could be resolved before the final vote. In total, 44 collaborators suggested 98 interventions, with each collaborator suggesting between one and six interventions ($M$ = 2.28, $SD$ = 1.35). The majority suggested changing the formulation of the base oath or adding information to it (80.6%), whereas 24.5% suggested adding an external factor (some interventions involved both; see Supplementary Note 3 for an overview of all suggestions). The three collaborators who eventually left the project all suggested interventions prior to leaving.

Suggested interventions were screened by the first author for potential duplicates. A total of 59 interventions grouped into 17 themes were identified as potentially overlapping and two independent coders rated the similarity and predicted effectiveness in reducing dishonesty within each theme. Based on these codings, 35 interventions were excluded as duplicates and 63 interventions were retained for the final vote (see Supplementary Notes 3 and 4 for detailed information on screening).

Collaborators were then invited to rate the retained 63 interventions based on (i) their perceived effectiveness compared to the control, measured on an 11-point scale from 0 (not at all) to 10 (extremely); (ii) their predicted effectiveness on a slider scale ranging from −27 percentage points to +27 percentage points compared to the control and adding information on the respective standardized effect (Cohen's $d$ ranging from −/+ 0.82); and (iii) the perceived feasibility of the intervention (i.e., how difficult it would be to implement the intervention in practice) on a scale from 0 (very difficult) to 10 (very easy). The maximum number of the slider scale was determined based on Pilot Study II, which found an average tax compliance of 73% in the control intervention (therefore allowing for a maximum increase of 27 percentage points; see Supplementary Note 9). Collaborators saw the 63 interventions in random order and anonymously rated each intervention on all three measures. A total of 38 collaborators completed ratings. The overall mean rating was 3.82

(*SD* = 2.56) for perceived effectiveness, *M* = 5.85 (*SD* = 3.28) for perceived feasibility. On average, raters expected an increase in *M* = 4.81 in percentage points (*SD* = 5.12), which translates to an average increase in Cohen's *d* of *M* = 0.15 (*SD* = 0.16). As registered, we computed a weighted index based on $\frac{2}{4}mean(effectiveness) + \frac{1}{4}mean(feasibility) - \frac{1}{4}sd(effectiveness)$ for each intervention. Interventions high on this index are considered subjectively effective; that is, raters showed reduced variation in their rated effectiveness, and considered them feasible to implement in practice. We ranked interventions on this index. A detailed overview of the rating procedure and results is provided in Supplementary Note 3. The 20 highest-ranking interventions were shared with all collaborators so they could comment on them and suggest changes. Collaborators reviewed interventions according to whether the final formulation and motivation were clear, whether they manipulated one factor (which was preferable), whether the formulation built on the base oath, and whether it was a duplicate. During this round, several collaborators noticed that external interventions (e.g., adding punishment to the design) were not manipulated in a full-factorial manner. That is, they were designed to be added to the base oath intervention but not to the control intervention, since this would have meant including two interventions for each external intervention. Out of *n* = 32 collaborators, a majority (84.4%) voted to exclude the external interventions from the final selection, which affected two interventions from the final selection. A second round of reviews among collaborators was solicited after suggested changes were implemented and the two external interventions were removed, which was followed by a discussion of which two of eight additional interventions that ranked high in the rating task would replace the removed interventions. A detailed overview of the review process and the implemented changes can be found in the Supplementary Notes 6 and 7. An overview of the final 20 interventions (as well as the control condition and baseline intervention) is provided in Table 1.

**Main Study**

After selecting the final 20 honesty oaths, we recruited UK and US participants for the main study, which comprised an online tax evasion game whose effectiveness we had validated in three pilot studies (Supplementary Notes 8–10).

**Sample Size Determination.** We aimed at recruiting 1,000 participants per intervention based on our available resources, for a total target sample size of 22,000 participants. A simulation indicated that this design is sensitive to effect sizes as small as a change of 3.5 percentage points (Cohen's $d \approx .11$) with 95% power when considering multiple comparisons (and the Benjamini-Hochberg correction) and the data distribution observed in the pilot studies (see Supplementary Note 11 for detailed power analysis).

**Participants.** We recruited a total of 23,327 participants via the crowdsourcing platform Prolific.com[66,67] Based on our preregistered exclusion criteria, we excluded participants who were younger than 18 years of age ($n = 0$), failed one of the comprehension checks ($n = 932$), who answered the survey faster than one third of the median response time ($n = 28$), who failed an attention check ($n = 141$), or who reported a higher income than their earned income ($n = 720$). In line with our preregistered plan, we stopped data collection because less than 10% of the sample were excluded (7.81%).

The final sample size consisted of 21,506 participants (12,069 female, 9,047 male, 295 nonbinary, 90 other, 5 undeclared) ranging from 18 to 99 years of age ($M = 39.3$, $SD = 13.3$). Based on our preregistered recruitment plan, we first collected data only from participants based in the UK. After two weeks we had recruited 15,078 participants (before exclusions) and, based on our preregistered plan, recruited the remaining share from the US. The final sample included 13,801 participants located in the UK and 7,705 located in the US ($n = 12,746$ UK nationals; $n = 7,502$ US nationals; $n = 1,248$ other or dual citizenship; for more detailed information on recruitment, including recruitment periods, see Supplementary

Note 12). Participants received a base payment of £0.60 and could earn a bonus payment ($range_{bonus}$ = £0.00 to £1.02; $M_{bonus}$ = £0.49, $SD_{bonus}$= £0.15). As participants on Prolific.com are always paid in British pounds, regardless of their location, the US-based sample received the same materials as the UK-based sample, except the reference to the charity was adjusted and the AI-generated honesty oath was slightly altered with regard to the charity.

**Procedure.** The study included a 22-between-subjects design, with participants being randomly allocated to one of the 22 interventions. An overview of the interventions and cell sizes is provided in Table 1. Participants completed a tax evasion game, which involved reporting income generated from a number sorting task (the even/odd task) that was subject to a tax of 35%. Before or after the sorting task, participants were asked to commit to an honesty oath (not in the control intervention). Once they had completed the game and the honesty oath, they reported their earned income from the sorting task, then provided additional measures and demographic information.

After receiving participant information and providing informed consent, participants read instructions on the even/odd task, a number sorting task in which participants sorted eight numbers into two boxes depending on whether the numbers were even or odd. They then saw an example of the task and completed three comprehension questions (see https://osf.io/g23b4 for details) and an attention check ("Attention Check! Please, click on Neither agree or disagree (2)." with three answer options). If participants answered a comprehension question incorrectly, they were shown the instructions a second time and had another opportunity to complete the comprehension items, which were always shown on the same page as the instructions. If participants failed to answer a comprehension question correctly a second time, the survey was terminated. Note that we only implemented the second round of comprehension checks after we had already tested 200 participants (this was not explicitly preregistered).

The tax evasion game consisted of four rounds of the even/odd task. In every round, four even numbers and four odd numbers between 0 and 99 were randomly generated, and participants had 30 seconds to sort them all. They advanced to the next round after sorting all numbers correctly or after 30s if they failed to do so. Participants earned £0.01 for every remaining second on the clock after each round. For example, if they completed a task in 10s they earned an income of £0.20 for that round. Participants' income was summed across the four rounds, so that they could technically earn between 0 and £1.20 (actual range: £0 to £1.02, $M$ = £0.49, $SD$ = £0.15). Participants saw their earned income after each round and income was added to the already existing earnings. The task was based on previous research using tax evasion games.[14]

In all interventions except the control condition, participants were provided with one of the 21 honesty oaths and asked to commit to it. For each participant it was randomly decided whether the honesty oath was presented before the even/odd task ($n$ = 10,225) or after, directly prior to tax reporting ($n$ = 10,328; placement of honesty oath) and whether participants could commit by checking a box next to the oath ($n$ = 10,234) or by retyping the oath in a text field ($n$ = 10,319; type of commitment). The oath was presented as a picture file so it could not be copied and pasted; participants had to manually type it. Because voluntary commitment is an important feature of successful honesty oaths,[14] participants could proceed without checking the box or retyping the statement. In total, 444 participants across all interventions (2.16%) did not commit to the honesty oath. These frequencies differed significantly across interventions, with the two conditions with the longest oaths showing higher frequencies of participants not committing (see Supplementary Note 14, Supplementary Table 9). A small number of participants ($n$ = 65) indicated problems seeing the honesty oath (e.g., due to slow network connection). Participants who did not commit to

the oath or who reported an error were not excluded from the main analyses, but sensitivity analyses excluding these did not reveal substantial differences (Supplementary Note 20).

After the tax evasion game, participants were asked to report the income they had generated (*actual income*). They were reminded of their actual income on the page before the income reporting page, but not on the income reporting page itself (based on results from Pilot Study III; Supplementary Note 10). They were also reminded that their income was subject to a 35% tax and that all taxes would be donated to the British or American Red Cross, depending on the sample (this information was also presented in the initial instructions). Based on previous research,[14,30] we chose a charity recipient in order to model how taxes redistribute resources and wealth within a society. Participants were also provided with a slider scale that showed how much taxes they would need to pay depending on the income they reported. The maximum was set to the actual generated income and the minimum to zero income, and the slider default was set to 50% of the actual generated income (the middle of the scale). Participants were then asked to enter their income (before taxes) in a text field (*reported income*). A bonus of $actual\ income - 35\% \times reported\ income$ was paid out after the participant completed the study. For instance, a participant generating an actual income of £1.00 and truthfully reporting this income received £0.65 as a bonus payment (1 − .35*1), whereas a participant generating an actual income of £1.00 but reporting £0 received a bonus payment of £1.00 (1 − .35*0). There was therefore an incentive to underreport the actual income in order to evade higher taxes and keep more of the income - as typical for tax evasion games.[30] After reporting their income, participants answered questions measuring their Honesty-Humility, general trust, socioeconomic status, and sleep quality, as well as demographic information including gender, age, and nationality. As a final step, they were debriefed.

**Measures**

***Tax Compliance.*** Tax compliance (our main measure of dishonesty) was measured via the tax evasion game using participants' actual and reported incomes. The ratio of reported to actual income specifies tax compliance. A tax compliance of 100% indicates a fully honest participant, whereas a tax compliance of 0% defines a fully dishonest participant.

***Time Spent.*** We recorded the time participants took to read the informed consent, to commit to the oath, and to report their final income.

***Honesty-Humility.*** Participants completed the 4-item Honesty-Humility scale on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree; McDonald's $\omega = 0.68$).[68]

***Trust.*** Participants answered three items from the Generalized Trust Scale ("Most people are basically honest," "Most people are trustworthy," "I am trustful") on a 5-point scale from 1 (strongly disagree) to 5 (strongly agree; $\omega = 0.82$).[69]

***Subjective Socioeconomic Status.*** Participants located themselves on a MacArthur ladder measuring subjective socioeconomic status according to where they think they stood compared to other people in their country (UK/US), from 1 (at the bottom in *X*) to 10 (at the top in *X*).

***Sleep Quality.*** Participants rated their sleep quality the previous night from 0 (terrible) to 5 (excellent). This item was added for a different project and no analyses were performed on it.

***Demographics.*** Participants reported their gender (female, male, nonbinary, other), age, and nationality (UK, US, other).

In a final step, we asked participants what they thought the purpose of the study was and whether they had participated in a similar study before.[70] The majority of participants reported that they had not participated in a similar study before (81.76%) and around half of all participants (53.35%) guessed the purpose of the study. Our preregistration included a measure on tax morale, but we did not include it in the final study due to time constraints.

**Forecasting Study**

Finally, we conducted a forecasting study asking collaborators, behavioral scientists not taking part in the project, laypeople, experts working in the finance and insurance industry, and LLMs to predict the outcome of the main study.

**Participants.** We recruited five populations: collaborators, external behavioral scientists, laypeople, experts working in finance and insurance, and LLMs. We invited all 43 collaborators (at the time), of whom 34 completed the forecasting.

Behavioral scientists not taking part in the project were recruited via personal contacts and mailing lists. We sampled 50 responses and excluded 21 due to failed comprehension checks or incompleteness. The final sample comprised 29 behavioral scientists across various fields (behavioral economics: $n = 4$, business: $n = 1$, cognitive psychology: $n = 3$, economic psychology: $n = 5$, judgement and decision making: $n = 2$, marketing: $n = 3$, organizational behavior: $n = 1$, moral/social psychology: $n = 9$, accounting: $n = 1$, behavioral law: $n = 1$), of whom 48.28% were self-identified experts on unethical behavior or corruption. The sample included six PhD students, four postdocs, eight assistant professors, six associate professors, and five full professors. (This sample was recruited after the main study had been analyzed, but the results were not publicly available at the time; see Supplementary Note 16).

We recruited laypeople on Prolific.com, excluding participants who indicated in the platform's online screening that they worked in the finance and insurance industry. We registered to recruit 75 participants per country (UK/US). Sample size determination was based on resource availability.[71] We sampled 154 participants and excluded participants who failed a comprehension check ($n = 4$), completed the study faster than one third of the median duration ($n = 3$), or indicated working in finance and insurance ($n = 6$). The latter were included in the sample of experts in the finance and insurance industry. We also moved 27 participants originally recruited to the industry experts sample who indicated working in an

industry other than finance and insurance to the laypeople sample. The final laypeople sample included 167 participants (87 male, 77 female, 3 nonbinary) ranging from 20 to 72 years of age ($M$ = 40.6, $SD$ = 11.9). Among those, 94 indicated UK nationality, 69 US nationality, and four other nationalities.

We recruited experts working in the finance and insurance industry using Prolific's internal screening. We also contacted tax organizations to circulate the survey among their members but did not receive a reply. We recruited 126 participants and excluded those who failed a comprehension check ($n$ = 1), completed the survey faster than one third of the median time ($n$ = 3), or indicated working in an industry other than finance and insurance ($n$ = 27). The latter participants were included in the laypeople sample. In total, we recruited 99 participants (65 male, 33 female, 1 other) ranging from 19 to 70 years of age ($M$ = 40, $SD$ = 10.7), working in the finance and insurance industry for a mean of 13.05 years ($SD$ = 10.68; range 1 to 45 years). The majority of the sample indicated UK nationality ($n$ = 92; US: $n$ = 2; other: $n$ = 5).

We also included three LLMs: ChatGPT4, Bard, and LLAMA2. We had originally registered to use ChatSonic instead of LLAMA2 but did not receive any meaningful responses using the standardized instructions.

**Procedure.** After providing informed consent, participants in all samples received the same information on the aim of the main study and the forecasting task, which consisted of rating 21 honesty oath formulations (including the base oath intervention) on their effectiveness compared to the control intervention (no honesty oath). The interventions were presented in random order and participants used the same slider scale that collaborators had used to vote at the start of the project. LLMs were given the same instructions and a list of all 21 interventions in fixed order (for detailed information on LLM instructions and output, see Supplementary Note 15). The external behavioral scientists sample reported their academic

positions, their specific field, and whether they had worked with research on unethical behavior or corruption in the past. Both laypeople and experts from the finance and insurance industry reported demographic information including gender, age, and nationality, as well as the industry they worked in and for how long.

The forecasting study was performed while data collection for the main study was already underway or finished (Supplementary Note 16). At the time no results had been published and the first author, who was the only one with access to the data, did not take part in the forecasting survey. We allowed participants from the main study to take part in the forecasting survey since they could not know the results, and because having taken part in the study could help them understand the design (and probably improve their predictions). We controlled for whether participants had taken part in the main study (laypeople: $n$ = 36 *yes* [21.56%], $n$ = 85 *maybe* [50.90%]; experts in finance and insurance industry: $n$ = 25 *yes* [25.25%], $n$ = 51 *maybe* [51.52%]). Results indicated no difference in predictions between participants who had taken part in the main study and those who had not (Supplementary Table 52).

**Measures**

***Predicted Effectiveness.*** All samples were asked to rate the effectiveness of the 21 interventions on a slider scale from –27 percentage points to +27 percentage points. These numbers were based on Pilot Study II, which found a tax compliance of 73% in the control intervention, meaning that the most effective intervention reducing all dishonesty would be an increase of 27 percentage points. The minimum value of the slider was chosen for symmetry.

***Demographics.*** Participants in the general population and expert sample reported their gender (female, male, nonbinary, other), age, nationality (UK, US, other), the industry they worked in (including 21 answer options, see https://osf.io/t3sm4/), how long they had

been working in the industry (in years), and whether they had participated in the main study (yes, maybe, no). Participants in the external behavioral scientist sample completed items on their academic position (PhD student, postdoc, assistant professor, associate professor, full professor, other), their specific field (open text response), and whether they had engaged in research focusing on unethical behavior or corruption (yes, no).

**Analytic Strategy**

The pilot studies suggested that the distribution of our dependent variable would contain a combination of a binomial distribution with participants having a tax compliance of 0 or 1 and a distribution of continuous scores between 0 and 1. Therefore, the dependent variable followed a distribution with upper and lower bounds, which is difficult to analyze with common linear models. We therefore analyzed the data using an ordered beta regression, which has been suggested as a valid alternative,[72] using the glmmTMB package.[73] For all analyses the alpha level was set to .05. For most models, we computed average marginal effects using the marginaleffects package.

We used R (version 4.2.1[74]) and the R-packages broom (version 1.0.4[75]), dplyr (version 1.1.1[76]), ggplot2 (version 3.4.4[77]), ggpubr (version 0.4.0[78]), glmmTMB (version 1.1.7[79]), janitor (version 2.1.0[80]), lme4 (version 1.1-32[81]), marginaleffects (version 0.14.0[82]), meta.shrinkage (version 0.1.4[83]), papaja (version 0.1.1[84]), purrr (version 1.0.1[85]), psych (version 2.2.5[86]), qualtRics (version 3.1.7[87]), sjPlot (version 2.8.14[88]), stringr (Version 1.5.0[89]), tidyverse (version 1.3.2[90]), and TOSTER (version 0.8.0[91]) for our analyses.

**Acknowledgements**

Panagiotis Mitkidis was supported by AUFF-E-2019-9-4 NOVA; Yuval Feldman was supported by ERC Grant number: 101054656 / Project acronym: VCOMP; Anna Z. Czarna was supported by Grant 2018/30/E/HS6/00863 from the National Science Center, Poland.

## References


1. Sommerstein, A. H. & Torrance, I. C. *Oaths and Swearing in Ancient Greece*. (De Gruyter, 2014). doi:10.1515/9783110227369.
2. Ostrom, E. Collective Action and the Evolution of Social Norms. *The Journal of Economic Perspectives* **14**, 137–58 (2000).
3. Bruin, B. Pledging integrity: Oaths as forms of business ethics management. *Journal of Business Ethics* **136**, 23–42 (2016).
4. Gächter, S. & Schultz, J. F. Intrinsic honesty and the prevalence of rule violations across societies. *Nature* **531**, 496–499 (2016).
5. Abeler, J., Nosenzo, D. & Raymond, C. Preferences for Truth-Telling. *Econometrica* **87**, 1115–1153 (2019).
6. Gerlach, P., Teodorescu, K. & Hertwig, R. The truth about lies: A meta-analysis on dishonest behavior. *Psychological Bulletin* **145**, 1–44 (2019).
7. Leib, M., Köbis, N., Soraperra, I., Weisel, O. & Shalvi, S. Collaborative dishonesty: A meta-analytic review. *Psychological Bulletin* **147**, 1241 (2021).
8. Jacobsen, C., Fosgaard, T. R. & Pascual-Ezama, D. Why Do We Lie? A Practical Guide to the Dishonesty Literature. *Journal of Economic Surveys* **32**, 357–87 (2018).
9. Bellé, N. & Cantarelli, P. What causes unethical behavior? A meta-analysis to set an agenda for public administration research. *Public Administration Review* **77**, 327–339 (2017).
10. Hertwig, R. & Mazar, N. Toward a taxonomy and review of honesty interventions. *Current Opinion in Psychology* 101410 (2022).
11. Zickfeld, J. *et al.* Committed Dishonesty: A Systematic Meta-Analysis of the Effect of Social Commitment on Dishonest Behavior. Preprint at https://doi.org/10.31234/osf.io/j47ng (2023).

12. Mazar, N., Amir, O. & Ariely, D. The dishonesty of honest people: A theory of self-concept maintenance. *Journal of marketing research* **45**, 633–644 (2008).
13. Barkan, R., Ayal, S. & Ariely, D. Ethical dissonance, justifications, and moral behavior. *Current Opinion in Psychology* **6**, 157–161 (2015).
14. Jacquemet, N., Luchini, S., Malézieux, A. & Shogren, J. F. Who'll stop lying under oath? Empirical evidence from tax evasion games. *European Economic Review* **124**, 103369 (2020).
15. Shalvi, S., Gino, F., Barkan, R. & Ayal, S. Self-serving justifications: Doing wrong and feeling moral. *Current Directions in Psychological Science* **24**, 125–130 (2015).
16. Koning, L., Junger, M. & van Hoof, J. Digital signatures: a tool to prevent and predict dishonesty? *Mind Soc* **19**, 257–285 (2020).
17. Kristal, A. S. *et al.* Signing at the beginning versus at the end does not decrease dishonesty. *Proceedings of the National Academy of Sciences* **117**, 7103–7107 (2020).
18. Chou, E. Y. What's in a name? The toll e-signatures take on individual honesty. *Journal of Experimental Social Psychology* **61**, 84–95 (2015).
19. Cagala, T., Glogowsky, U., Rincke, J. & Schudy, S. Commitment requests do not affect truth-telling in laboratory and online experiments. *Games and Economic Behavior* **143**, 179–190 (2024).
20. Cagala, T., Glogowsky, U. & Rincke, J. Detecting and preventing cheating in exams: Evidence from a field experiment. *Journal of Human Resources* (2021).
21. Kettle, S., Hernandez, M., Sanders, M., Hauser, O. & Ruda, S. Failure to captcha attention: null results from an honesty priming experiment in guatemala. *Behav. Sci* **7**, 28 (2017).
22. Koessler, A.-K., Torgler, B., Feld, L. P. & Frey, B. S. Commitment to pay taxes: Results from field and laboratory experiments. *European Economic Review* **115**, 78–98 (2019).

23. Martuza, J. B., Skard, S. R., Løvlie, L. & Thorbjørnsen, H. Do honesty-nudges really work? A large-scale field experiment in an insurance context. *Journal of Consumer Behaviour* **21**, 927–951 (2022).
24. Retraction for Shu et al., Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proc Natl Acad Sci U S A* **118**, e2115397118 (2021).
25. Peer, E., Mazar, N., Feldman, Y. & Ariely, D. Honesty Pledges: the Effects of Involvement and Identification Over Time. *Available at SSRN 4355553* (2023).
26. Balafoutas, L., Beck, A., Kerschbamer, R. & Sutter, M. The hidden costs of tax evasion.: Collaborative tax evasion in markets for expert services. *Journal of Public Economics* **129**, 14–25 (2015).
27. Duckworth, A. L. & Milkman, K. L. A guide to megastudies. *PNAS Nexus* **1**, pgac214 (2022).
28. Milkman, K. L. *et al.* Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
29. Milkman, K. L. *et al.* A 680,000-person megastudy of nudges to encourage vaccination in pharmacies. *Proceedings of the National Academy of Sciences* **119**, e2115126119 (2022).
30. Alm, J. & Malézieux, A. 40 years of tax evasion games: a meta-analysis. *Exp Econ* **24**, 699–750 (2021).
31. Alm, J., Bloomquist, K. M. & McKee, M. On the External Validity of Laboratory Tax Compliance Experiments. *Economic Inquiry* **53**, 1170–1186 (2015).
32. Skowronek, S. E. DENIAL: A Conceptual Framework to Improve Honesty Nudges. *Current Opinion in Psychology* 101456 (2022) doi:10.1016/j.copsyc.2022.101456.

33. Gollwitzer, P. M. & Sheeran, P. Implementation intentions and goal achievement: A meta-analysis of effects and processes. *Advances in experimental social psychology* **38**, 69–119 (2006).
34. Mitkidis, P., Sørensen, J., Nielbo, K. L., Andersen, M. & Lienard, P. Collective-Goal Ascription Increases Cooperation in Humans. *PLOS ONE* **8**, e64776 (2013).
35. Mulder, L. B., Jordan, J. & Rink, F. The effect of specific and general rules on ethical decisions. *Organizational Behavior and Human Decision Processes* **126**, 115–129 (2015).
36. Mulder, L. B., Rink, F. & Jordan, J. Constraining temptation: How specific and general rules mitigate the effect of personal gain on unethical behavior. *Journal of Economic Psychology* **76**, 102242 (2020).
37. Schild, C., Heck, D. W., Scigala, K. A. & Zettler, I. Revisiting REVISE: (Re)Testing unique and combined effects of REminding, VIsibility, and SElf-engagement manipulations on cheating behavior. *Journal of Economic Psychology* **75**, 102161 (2020).
38. Verschuere, B. *et al.* Registered Replication Report on Mazar, Amir, and Ariely (2008). *Advances in Methods and Practices in Psychological Science* **1**, 299–317 (2018).
39. Zhao, J., Dong, Z. & Yu, R. Don't remind me: When explicit and implicit moral reminders enhance dishonesty. *Journal of Experimental Social Psychology* **85**, 103895 (2019).
40. Toor, N. S. Comparison of Dishonesty Interventions: A Conceptual Replication Study. (2022).
41. Jamison, J. C., Mazar, N. & Sen, I. Applying behavioral insights to tax compliance: experimental evidence from Latvia. (2021).

42. Ayal, S., Celse, J. & Hochman, G. Crafting messages to fight dishonesty: A field investigation of the effects of social norms and watching eye cues on fare evasion. *Organizational Behavior and Human …* (2019).
43. Lois, G. & Wessa, M. Honest mistake or perhaps not: The role of descriptive and injunctive norms on the magnitude of dishonesty. *Journal of Behavioral Decision Making* **34**, 20–34 (2021).
44. Hallsworth, M., List, J. A., Metcalfe, R. D. & Vlaev, I. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics* **148**, 14–31 (2017).
45. Brudermann, T., Bartel, G., Fenzl, T. & Seebauer, S. Eyes on social norms: A field study on an honor system for newspaper sale. *Theory Decis* **79**, 285–306 (2015).
46. Köbis, N. C., Troost, M., Brandt, C. O. & Soraperra, I. Social norms of corruption in the field: social nudges on posters can help to reduce bribery. *Behavioural Public Policy* **6**, 597–624 (2022).
47. Fellner, G., Sausgruber, R. & Traxler, C. Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information. *Journal of the European Economic Association* **11**, 634–660 (2013).
48. Dimant, E., Van Kleef, G. A. & Shalvi, S. Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior & Organization* **172**, 247–266 (2020).
49. Castro, L. & Scartascini, C. Tax compliance and enforcement in the pampas evidence from a field experiment. *Journal of Economic Behavior & Organization* **116**, 65–82 (2015).
50. Köbis, N., Starke, C. & Rahwan, I. The promise and perils of using artificial intelligence to fight corruption. *Nat Mach Intell* **4**, 418–424 (2022).

51. Capraro, V. *et al.* The impact of generative artificial intelligence on socioeconomic inequalities and policy making. Preprint at http://arxiv.org/abs/2401.05377 (2023).
52. Guzikevits, M. & Choshen-Hillel, S. The optics of lying: How pursuing an honest social image shapes dishonest behavior. *Current Opinion in Psychology* **46**, 101384 (2022).
53. Alm, J. & Torgler, B. Culture differences and tax morale in the United States and in Europe. *Journal of economic psychology* **27**, 224–246 (2006).
54. Heck, D. W., Thielmann, I., Moshagen, M. & Hilbig, B. E. Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making* **13**, 356–371 (2018).
55. DellaVigna, S. & Pope, D. Predicting Experimental Results: Who Knows What? *Journal of Political Economy* **126**, 2410–2456 (2018).
56. Zickfeld, J. H., Ściga\la, K. A., Weiss, A., Michael, J. & Mitkidis, P. Commitment to honesty oaths decreases dishonesty, but commitment to another individual does not affect dishonesty. *Communications Psychology* **1**, 27 (2023).
57. Götz, F. M., Gosling, S. D. & Rentfrow, P. J. Small Effects: The Indispensable Foundation for a Cumulative Psychological Science. *Perspect Psychol Sci* **17**, 205–215 (2022).
58. Jacquemet, N., Luchini, S., Rosaz, J. & Shogren, J. F. Truth- telling under oath. *Management Science* (2018).
59. Peer, E. & Feldman, Y. Honesty pledges for the behaviorally-based regulation of dishonesty. *Journal of European Public Policy* **28**, 761–781 (2021).
60. Jacquemet, N., James, A. G., Luchini, S., Murphy, J. J. & Shogren, J. F. Do truth-telling oaths improve honesty in crowd-working? *PloS one* **16**, e0244958 (2021).
61. Jacquemet, N., Luchini, S., Rosaz, J. & Shogren, J. F. Can We Commit Future Managers to Honesty? *Frontiers in Psychology* **12**, (2021).

62. Jacquemet, N., Joule, R.-V., Luchini, S. & Shogren, J. F. Preference elicitation under oath. *Journal of Environmental Economics and Management* **65**, 110–132 (2013).
63. Jacquemet, N., James, A., Luchini, S. & Shogren, J. F. Referenda under oath. *Environmental and resource economics* **67**, 479–504 (2017).
64. Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E. & Thies, W. Deterring Cheating in Online Environments. *ACM Trans. Comput.-Hum. Interact.* **22**, 28:1-28:23 (2015).
65. Uhlmann, E. L. *et al.* Scientific utopia III: Crowdsourcing science. *Perspectives on Psychological Science* **14**, 711–733 (2019).
66. Palan, S. & Schitter, C. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* **17**, 22–27 (2018).
67. Peer, E., Rothschild, D., Gordon, A., Evernden, Z. & Damer, E. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* **54**, 1643–1662 (2022).
68. De Vries, R. E. The 24-item brief HEXACO inventory (BHI). *Journal of Research in Personality* **47**, 871–880 (2013).
69. Yamagishi, T. & Yamagishi, M. Trust and commitment in the United States and Japan. *Motivation and emotion* **18**, 129–166 (1994).
70. Skowronek, S. About 70% Of Participants Know That The Canonical Deception Paradigms Measure Dishonesty. in *Academy of Management Proceedings* vol. 2021 13725 (Academy of Management Briarcliff Manor, NY 10510, 2021).
71. Lakens, D. Sample size justification. *Collabra: Psychology* **8**, 33267 (2022).
72. Kubinec, R. Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. *Political Analysis* **31**, 519–536 (2023).
73. Bolker, B. Getting started with the glmmTMB package. *Vienna, Austria: R Foundation for Statistical Computing. software* (2016).

74. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2022).
75. Robinson, D., Hayes, A. & Couch, S. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom (2023).
76. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr (2023).
77. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
78. Kassambara, A. *Ggpubr: 'ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr (2020).
79. Brooks, M. E. *et al.* glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal* **9**, 378–400 (2017).
80. Firke, S. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor (2021).
81. Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).
82. Arel-Bundock, V. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects (2023).
83. Taketomi, N. & Emura, T. *Meta.Shrinkage: Meta-Analyses for Simultaneously Estimating Individual Means*. https://CRAN.R-project.org/package=meta.shrinkage (2023).
84. Aust, F. & Barth, M. *papaja: Prepare Reproducible APA Journal Articles with R Markdown*. https://github.com/crsh/papaja (2022).
85. Wickham, H. & Henry, L. *Purrr: Functional Programming Tools*. https://CRAN.R-project.org/package=purrr (2023).

86. Revelle, W. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. https://CRAN.R-project.org/package=psych (2022).
87. Ginn, J., O'Brien, J. & Silge, J. *qualtRics: Download 'qualtrics' Survey Data*. https://CRAN.R-project.org/package=qualtRics (2022).
88. Lüdecke, D. *sjPlot: Data Visualization for Statistics in Social Science*. https://CRAN.R-project.org/package=sjPlot (2023).
89. Wickham, H. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr (2022).
90. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of open source software* **4**, 1686 (2019).
91. Lakens, D. Equivalence Testing With TOSTER. *APS Observer* **30**, (2017).
92. Schweitzer, M. E. & Hsee, C. K. Stretching the truth: Elastic justification and motivated communication of uncertain information. *Journal of Risk and Uncertainty* **25**, 185–201 (2002).
93. Hernandez, M., Jamison, J., Korczyc, E., Mazar, N. & Sormani, R. Applying behavioral insights to improve tax collection. (2017).
94. Capraro, V. Gender differences in lying in sender-receiver games: A meta-analysis. *Judgment and Decision making* **13**, 345–355 (2018).