Do Scientists Tell the Truth?  
Evidence from a Field Experiment

Moritz A. Drupp\textsuperscript{a}, Menusch Khadjavi\textsuperscript{a, b} and Rudi Voss\textsuperscript{a, *}, †

\textsuperscript{a} Department of Economics, University of Kiel, Germany  
\textsuperscript{b} Kiel Institute of the World Economy, Germany

This version: April 18, 2018

Academic honesty is crucial for the advancement of and trust in science. However, survey evidence suggests that a considerable number of scientists engage in questionable research practices. Motivated by identity economics theory, we provide evidence on incentivized truth-telling behavior of scientists by means of an online field experiment. We conduct an established coin-tossing task with 437 members of an international scientific organization, in which participants face a trade-off between monetary incentives of lying and honest reporting. In particular, we compare reporting behavior across two treatments, either making the private or professional identity more salient. We find that fewer scientists over-report winning tail tosses in the professional identity treatment. Furthermore, we find that a number of measures of scientific output are associated with truth-telling. Yet we also find that scientists over-report tail tosses compared to the truthful distribution even in the professional identity treatment. While honesty norms associated with the scientific identity thus seem to increase truth-telling, academia still has to further foster norms of honest behavior and enforce measures for preventing scientific misbehavior.

\textbf{JEL codes:} C93, D82, K42, J45  
\textbf{Keywords:} Truth-telling, lying, professional identity, science, field experiment

\textsuperscript{*} Corresponding author: Moritz A. Drupp, Department of Economics, University of Kiel, Wilhelm-Seelig-Platz 1, 24118 Kiel, Germany. Email: Drupp@economics.uni-kiel.de.

\textsuperscript{†} We are grateful to Alain Cohn, Sarah Necker, Martin Quaas, Marie-Catherine Riekhof and Jörn Schmidt as well as audiences at UCSB and UCSD for helpful discussions. We thank the participating scientists, the administrative office of the science organization for providing e-mail address data, Jörn Schmidt for handling correspondences, Marie-Catherine Riekhof for helping with the study design, Olaf Bock and his team at the Experimental Lab in Hamburg for help in administering the experiment, as well as Pia Förster, Clara Paczkowski and Lasse Renz-Kiefel for research assistance. This work was supported by the German Ministry of Education and Research [grant 01UT1410] and by the Cluster of Excellence 80, which is funded by the German Research Foundation (DFG).
1. Introduction

Whether and to what degree scientists behave ethically sound and tell the truth is of fundamental importance for the development of science, for public trust in science and, as such, for the future of mankind. Marshall (2000: 1162) called this “a Million-Dollar Question”, but this number is likely a gross underestimate. This is particularly true for times which call, on the one hand, for more ‘evidence-based policy-making’ and are otherwise guided by low trust in scientists and a tendency to blur distinctions between objective knowledge and so-called ‘alternative facts’, ‘fake news’ and ‘post-truths’.

The Merriam-Webster Dictionary (2017) defines science as “knowledge or a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through scientific method”. The quest for ensuring integrity in research conduct is probably as old as science itself, yet the reputation of truthful science has in particular suffered in recent times from prominent instances of scientific misconduct. A famous, now retracted article by Wakefield et al. (1998) suggested that vaccinating children against measles, mumps and rubella increases the risk of autism. Poland and Jacobson (2011) describe the public reaction of anti-vaccination campaigns to the now disproved article. In the time following the publication of Wakefield et al. (1998), there was a record of hundreds of cases of measles outbreaks and even some children dying (Poland and Jacobson, 2011), providing some indication of the tremendous social costs of scientific misconduct.

Beyond the prominent cases of scientific misconduct mentioned above, survey evidence suggests that a considerable number of scientists engage in a broader set of questionable research practices (see, for example, John et al., 2012; List et al., 2001; Martinson et al., 2005; Necker, 2014). A meta-study by Fanelli (2009) summarizes findings from 21 individual studies and shows that around two percent of scientists admit to having committed serious forms of scientific misconduct at least once, such as fabricating, falsifying or modifying data or results. It further supports findings of a

---

1 In fact, ‘post-truth’ was the word of the year 2016, as elected by Oxford Dictionaries (see BBC, 2016).

2 These include, among others, cases such as of the cloning expert Hwang Woo-suk, the evolutionary biologist Marc Hauser, and social psychologist Diede Stapel. Articles by Sang-Hun (2009), Wade (2010) and Bhattacharjee (2013) provide more information on the respective misconduct.

3 Besides anonymous survey-based approaches, there exist a number of other recent examples testing research integrity and the robustness of scientific research: For example, Camerer et al. (2016) replicated eighteen recent prominent experimental economics works. They find that about two-thirds of all findings can be replicated; Brodeur et al. (2016) provide recent evidence that the reporting of empirical findings tends to be biased towards regression specifications that favor rejecting the null hypothesis. In order to improve research practices, Simmons et al. (2011) recently proposed rules of sound scientific conduct in order to decrease so-called experimenter degrees of freedom.
previous study by Martinson et al. (2005), showing that as many as one-third of scientists admit to have engaged in questionable research practices, such as ‘using another’s ideas without obtaining permission or giving due credit’, ‘failing to present data that contradict one’s own previous research’, or ‘inappropriately assigning authorship credit’. This literature suggests that the search for general truths is not always conducted in a truthful manner. Yet, this evidence so far only relies on anonymous survey responses, with the fundamental challenge that there is no individual (monetary) incentive to participate and to report truthfully.

Our study provides experimental economic evidence on incentivized truth-telling of more than 400 scientists by means of an online field experiment. We thus provide evidence that can be viewed as complementary to above mentioned survey approaches. Specifically, our aim is to investigate whether the professional identity as a scientist affects honesty, i.e. whether the professional identity as a scientist motivates and fosters truthful behavior. After all, science ‘consists in the search for truth’ (Popper, 1996).

To this end, we employ a simple coin-tossing task in which scientists are asked to toss a fair coin four times and report back their number of tail tosses (Abeler et al., 2014). For each reported tail toss, they receive five Euros. While individual (dis)honesty is not detectable, we can estimate the deviation of reported tosses against the expected truthful distribution. Studying individuals’ truth-telling in this manner has become a major research focus in economics. Furthermore, this measure has been shown to carry external validity as it correlates with truth-telling behavior beyond the simple experimental task.

To study whether professional identity of scientists induces more honesty, we draw on the identity priming literature that was developed in social psychology and is now an active research field within economics (see Cohn and Maréchal (2016) for a

4 We thereby also contribute to the still rather scarce literature on the behavioral economics of science and academia. Among others, Gächter et al. (2009) study how framing impacts the decision to choose when to register for an academic conference, Löfgren et al. (2012) scrutinize the impact of a default option on uptake of carbon offsetting among environmental economists, and Chetty et al. (2014) conduct an experiment on pro-social behavior with referees of the Journal of Public Economics.

5 Recently, two studies have examined how the professional identity of participants and associated norms affect truth-telling behavior. Cohn et al. (2014, 2015) provide experimental evidence that bankers and prisoners behave less honestly when their respective professional identity is made salient as compared to a (private) control identity (cf. Villeval, 2014).


7 See, for instance, Cohn et al. (2015), Cohn and Maréchal (2018), Dai et al. (2018), Drupp et al. (2016), Gächter and Schulz (2016), as well as Potters and Stoop (2016).
The idea is that individuals have multiple identities that are guided by different norms and behavioral patterns (Akerlof and Kranton, 2000). Individuals experience disutility if they deviate from norms prescribed by their respective salient identity. Our experiment accordingly consists of two treatments. The professional identity treatment aims at making the subject’s professional identity as a scientist salient, while the private identity (control) treatment aims at making the private identity salient. To prime participants, we use nine questions that are designed to capture common features of a professional or private context, that are unrelated to truth-telling and as similar as possible across the two treatments. For example, subjects in the professional identity treatment were asked “Where did you last go to for a conference/workshop?” and “What activity in your work do you enjoy the most?”, while subjects in the private identity (control) treatment were asked “Where did you last go on holiday?” and “What activity in your leisure time do you enjoy the most?”. In the context of our study, the priming intervention aims to reveal the behavioral difference between a subject’s private and professional identity and thus be indicative of the norms and behavioral patterns associated with the scientific identity of the subjects in terms of truth-telling and honesty.

Based on 437 responses to our coin-tossing task, we find that significantly fewer scientists over-report winning coin tosses in the professional identity treatment compared to the private identity control. The scientific identity therefore seems to entail stronger honesty norms that induce more frequent truth-telling. Nevertheless we find that even those who were primed in their professional identity significantly over-report winning tail tosses compared to the expected truthful frequency. While honesty norms associated with scientific identity thus already serve to increase truth-telling, science still has to establish rigorous measures for preventing scientific misbehavior to ensure that science is not derailed from its path to generate truths.

---

8 As our study concerns the technique of priming and focuses on truth-telling behavior, it is worthwhile to note that there are doubts about the robustness of results obtained in the priming literature in social psychology and suspicions that questionable research practices have been employed. As a response to this critique, Daniel Kahneman called for systematic replication efforts in this field (Young, 2012). Not specifically scrutinizing priming studies, Camerer et al. (2016) and Open Science Collaboration (2015) have recently demonstrated that such large-scale replication attempts are feasible and fruitful.
To study the truth-telling of scientists, we conducted an online field experiment with members of an international scientific organization that was established more than 100 years ago. The administrative office of the organization provided an e-mail list of its 1,930 members. In the summer of 2016 we contacted all members by e-mail and invited them to participate in a short online study that consisted of ten pages and took about 15 minutes to complete. We told them that they could earn 25 € on average (equivalent to $27 at the time of the experiment) for participating, with the exact individual earnings depending on chance and their choices. We ensured that their individual responses are kept confidential and informed the participants about the confidentiality.

Upon clicking the link to the online study in the invitation e-mail, subjects were assigned to one of two treatments by the online platform: either the professional identity treatment (abbreviated Professional or PROF) or the private identity (control) treatment (Private or PRIV). A preamble page provided further details on the experiment and the mode of payment (Amazon vouchers). The study then began with simple descriptive questions on age, gender and nationality. This was followed by our manipulation that consisted of nine questions either relating to their professional identity (Professional treatment) or relating to their private identity (Private treatment). The purpose of these questions was to make the subjects’ professional identity as scientists, and associated norms, more salient in Professional as compared to Private.

The behavioral intervention of identity priming builds on a by now established strain of the experimental economics literature. The basic idea—based on Akerlof and Kranton (2000)—is that people have multiple identities that are guided by different norms and behavioral patterns. Individuals experience disutility if they deviate from norms prescribed by their respective salient identity. This depends on the relative weight of that identity. The technique of identity priming aims at making a given identity, such as the professional identity of being a scientist, temporarily more salient (see, e.g., Benjamin et al., 2010, 2016; Cohn and Maréchal, 2016, Cohn et al., 2014, 2015, 2018).

The members are predominantly natural scientists, with a focus on the marine environment. We do not report the name of the scientific organization to increase the anonymity of our respondents.

Cohn and Maréchal (2016) provide a review of identity priming in economics and discuss how this builds on a previous substantial literature in social psychology. The first economic experiments on identity priming were Chen and Li (2009) as well as Benjamin et al. (2010). There are two general approaches to studying how behavioral measures differ across identities: (1) artificially inducing certain identities or (2) studying the effect of identity priming in natural populations, such as bankers (Cohn et al., 2014), criminals, (Cohn et al., 2015), or scientists, as in our study.
Our study design closely builds on the approach of Cohn et al. (2014, 2018). The priming intervention should reveal the behavioral difference between a subject’s private and professional identity. Thus, the intervention should be indicative of the norms and behavior associated with the scientific identity as compared to the private identity of the subjects in terms of truth-telling and honesty. In an effort to reduce potential confounding due to priming effects that are unrelated to their private or professional identity, we designed the questions to capture salient features of their professional work or private life identity, yet to be as similar as possible in terms of their content and context. For example, subjects in the professional treatment were asked “Where did you last go to for a conference/workshop?”, while subjects in the private control treatment were asked “Where did you last go on holiday?” (see Table 1 for a list of all priming questions posed and Appendix A for screenshots from the online survey). These priming questions were the only difference between the two treatment conditions.\(^{11}\)

### Table 1: Identity priming questions

<table>
<thead>
<tr>
<th>Professional identity treatment</th>
<th>Private identity treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who is your current employer?</td>
<td>What is your current city of residence?</td>
</tr>
<tr>
<td>How many years have you worked for this institution?</td>
<td>How many years have you lived in your current accommodation?</td>
</tr>
<tr>
<td>Do you have a tenured position?</td>
<td>Are you married?</td>
</tr>
<tr>
<td>How large is your direct working team (yourself included)?</td>
<td>How large is your direct family (yourself included)?</td>
</tr>
<tr>
<td>Where did you last go to for a conference/workshop?</td>
<td>Where did you last go on holiday?</td>
</tr>
<tr>
<td>In which year did you start your PhD?</td>
<td>In which year did you kiss the first boy/girl?</td>
</tr>
<tr>
<td>At what time do you usually arrive at the office?</td>
<td>At what time do you usually arrive at home?</td>
</tr>
<tr>
<td>What activity in your work do you enjoy the most?</td>
<td>What activity in your leisure time do you enjoy the most?</td>
</tr>
<tr>
<td>How satisfied are you with your work in general?</td>
<td>How satisfied are you with your life in general?</td>
</tr>
</tbody>
</table>

\(^{11}\) The only other difference was that on the preamble page we stated that the study was on either on “Work [Life] satisfaction, including individual attitudes and behavior” in Professional [Private].
This identity manipulation was followed by three experimental tasks. First, subjects were asked to complete a risk preference elicitation task based on Binswanger (1981) and Eckel and Grossman (2002), the results of which we analyze in a companion paper (Drupp et al., 2017). The risk task was followed by the truth-telling task based on Abeler et al. (2014) that is the main focus of this paper. We present this task in more detail below. Finally, we posed a hypothetical social time preference task. The three tasks were always presented in this order and it was not possible to switch back once a subject had proceeded to the next page. The lottery outcome of the risk task was only revealed at the end of the experiment and thus could not have affected coin toss reporting.

Following the experimental tasks, participants were also asked to complete a short follow-up survey that included a word-completion task designed to provide an implicit measure of how well the identity priming manipulation had worked (cf. Cohn et al., 2014). Subjects were presented with eight word fragments and they were asked to fill in the gaps with letters to form existing words. The idea is that when the professional identity is salient other words come to the participants’ mind as compared to when the private identity is salient. For example, they were shown the word fragment “j o u r_ _ _”, which they could complete with the word “journal” that scientists would frequently encounter in their professional lives, or the word “journey,” which might be more salient to those in the Private treatment.\(^{12}\) We classified all completed words and either assigned the number 1 to words related to the professional work identity or number 0 to words classified as related to a private life. Words that could not be classified as relating to either context or words without actual meaning were coded as missing.\(^ {13}\)

Together with the payoff from the risk elicitation task, ranging from 2 to 16€, and a 5€ compensation for completing the short follow-up survey, each subject could earn up to 41€.\(^ {14}\) The payoff from the risk task was revealed after subjects had completed the follow-up survey. Finally, we offered the possibility to donate fractions of the earnings to the charity ‘Doctors Without Borders’.\(^ {15}\) Subsequent to the online experiment in May and June 2017, we also collected observable information of the participating scientists in the database SCOPUS, including the number of publications and citations.

---

\(^ {12}\) The first two of the eight word fragments (“_ a l k” and “_ o o k”) had no unambiguous professional science interpretation. These two were meant as an easy start for participants and served, following Cohn et al. (2014, 2017), the purpose of disguising the purpose of the task. The other word fragments were: “_ i s _”, “_ s s i o n”, “c o _”, “_ o c k” as well as “_ p e r”.

\(^ {13}\) When in doubt about a word’s meaning we relied on the Merriam-Webster dictionary.

\(^ {14}\) The design thus aimed at paying out all participants. Overall, we spent 3,389 Euros on subject remuneration and donated 6,199 Euros to ‘Doctors Without Borders’ on our participants’ behalf.

\(^ {15}\) This donation option was not pre-announced and it thus could not have influenced coin toss reporting.
For studying the truth-telling of scientists, we adapt the 4-coin-tossing task of Abeler et al. (2014) for our online field experiment. Subjects were asked to use any coin that has the usual “tails” and “heads” format (see Appendix A for a screenshot of the task). The subject’s task was then to toss this coin exactly 4 times, and report their tail toss result by clicking on the relevant button in a table. For each instance they reported that the winning toss “tails” laid on top, they received 5 €. An important feature of this task is that lying can be detected only on aggregate when examining the distribution of decisions, but not on the individual level. Thus, depending on chance and honesty, each subject received between 0 and 20 € for this task. Similar experiments using coin tosses or die rolling have been conducted to answer a whole range of related research question. Abeler et al. (2016) provide a meta-study on truth-telling behavior summarizing results based on 72 individual studies. Several key insights emerge from this burgeoning literature: (i) Subjects only over-report on average a quarter of the possible maximum pay-off and thus exhibit substantial lying costs; (ii) Subject’s reporting behavior is not significantly influenced by stake sizes; (iii) female subjects over-report somewhat less compared to males; (iv) students over-report more than other subjects. Testing different models that can be used to explain reporting behavior, Abeler et al. (2016) find that models that combine a preference for being honest, i.e. that entail a utility cost for deviating from the truthful response, and preference for being seen as honest, i.e. that entail individual reputation concerns, perform best in explaining experimental data.

As our main contribution is not a focus on modeling lying costs but more directly on the effect of making the professional scientific identity more salient vis-a-vis the private identity, we follow Benjamin et al. (2010) and Cohn et al. (2015) in relying on a simple behavioral choice model that features the salience of distinct identities. The model of reporting behavior considers an overall lying aversion due to deviating from the truthful response that may differ between the two identities, which may be guided by different norms and behavioral patterns.

16 As we could not ensure the availability of coins to toss remotely, we offered the option of to proceed without reporting one of the five tail toss possibilities in case they could not organize a coin to toss. They were told that they would not receive a payoff for this task in this case. No subject clicked this option.

17 Another recent study by Gneezy et al. (2018) investigates how lying costs depend on the size of the lie in various dimensions using both unobservable as well as observable lying tasks. Besides intrinsic lying costs considered in our model, they find that an important role for reputational concerns driving honest reporting in unobservable games, such as our coin tossing experiment. Furthermore, they find that only one out of 602 subjects under-reports to his or her disadvantage.

18 Besides the application of identity-priming model to truth-telling behavior of criminals by Cohn et al. (2015), this model has been employed for explaining effects of religious identity on a suite of economic preferences (Benjamin et al., 2016) and on risk preferences (Cohn et al., 2017; Drupp et al., 2017).
In absence of a possibility to detect individual lying, an individual $i$ faces a trade-off between monetary incentives and (moral) costs of lying. While the individual derives utility only from her payoff proportional to the reported number of coin tosses $r_i$, she also suffers disutility from reporting a number that deviates from the true number of tail tosses, $r_{it}$. The individual payoff-maximizing choice is given by $r_{ip}$. Aggregating over all $n$ individuals of a population yields the mean tail toss reporting $\bar{R} = \frac{1}{n}\sum_{i=1}^{n} r_i$, which can be disaggregated for different groups within a population. For instance, we denote the mean tail toss reporting in the Professional/identity treatment as $\bar{R}^{PROF}$.

Furthermore, let $\hat{R}^{PROF}(\hat{R}^{PRIV})$ denote the expected reporting behavior implied by prevailing norms in the professional environment (private identity context). In the context of our study, these norms imply certain lying costs, with $\hat{R} = \frac{\lambda}{2} (r_i - r_{it})$, where $\lambda$ is a parameter determining the degree of overall lying aversion. As the degree of lying aversion may depend on expected behavior and prevailing norms in different contexts, it may in particular differ across the private and the professional identity conditions, i.e. $\lambda^{PROF} \neq \lambda^{PRIV}$ and thus $\hat{R}^{PROF} \neq \hat{R}^{PRIV}$. Furthermore, let $s$ denote the strength of the identification with the professional environment. Let $w_i(s) \in [0,1]$ denote how much weight the individual puts on complying with expectations in the professional environment, which depends on the strength of identifying with the respective environment, with $\frac{\partial w_i}{\partial s} \geq 0$. In this set-up, the individual chooses her reporting $r_i$ to maximize utility

$$\max_{r_i} U_i(r_i) = -\frac{1}{2} (1 - w_i(s))(r_i - \hat{R}^{PRIV})^2 - \frac{1}{2} w_i(s)(r_i - \hat{R}^{PROF})^2. \quad (1)$$

The optimal tail toss reporting $r_i^*$ is a weighted average of the ‘expected’ reportings under both identities,

$$r_i^* = (1 - w_i(s))\hat{R}^{PRIV} + w_i(s)\hat{R}^{PROF}. \quad (2)$$

In terms of the model, our priming experiment aims at varying the salience of the professional or the private identity and thus the strength $s$ of identifying with the professional identity. Priming subjects with the professional identity (the Professional treatment) should increase $s$, while priming the private identity (the Private treatment) should decrease $s$. Subjects should therefore (weakly) experience an increase in the

---

19 While the model considers continuous reporting, our subsequent experiment is based on a setting where possible reporting levels are discrete, with $r_i, r_{it} \in \{0,4\}$. Furthermore, the mean truthful response is given by $R_t = \frac{1}{n}\sum_{i=1}^{n} r_{it} = 2$, and the payoff-maximizing choice is given by $R_p = \frac{1}{n}\sum_{i=1}^{n} r_{ip} = 4$. 
weight put on one identity or the other when completing our experimental task. As such, the treatment effect should reveal the marginal behavioral impact of the primed identity and its associated norms relative to the other treatment,

$$\frac{\partial r^*_i}{\partial s} = \frac{\partial w_l}{\partial s} (\hat{R}^{PROF} - \hat{R}^{PRIV})$$

(3)

Based on previous findings in the experimental literature (Abeler et al., 2016), we expect heterogeneity regarding individual truth-telling $r^*_i$ in our sample of scientists. Translating the average standardized estimate of the meta-study of Abeler et al. (2016) into our context predicts an average tail toss report $\bar{R}$ of 2.44. We formulate:

**Hypothesis 1:** Average over-reporting is in-between the truthful and the payoff maximizing choice.

While previous research has shown that professional identity is associated with higher over-reporting of winning coin tosses (i.e. lower truth-telling) for bankers and criminals (Cohn et al., 2014, 2015), we hypothesize that the norms and behavioral patterns associated with working as a scientists implies greater truth-telling. After all, science is a system of knowledge covering general truths (Popper, 1996). We therefore assume greater lying costs in the professional science context, $\lambda^{PROF} > \lambda^{PRIV}$, and accordingly norms associated with lower expected mean tail toss reporting, $\hat{R}^{PROF} < \hat{R}^{PRIV}$. Our model thus predicts that $\frac{\partial r^*_i}{\partial s} < 0$, summarized as

**Hypothesis 2:** Average over-reporting of scientists is lower in the professional identity treatment.

Even though we expect that stronger honesty norms are present in the professional scientific as compared to the average private context, the accumulating evidence on the use of questionable research practices among scientists suggests that we should not expect truthful reporting on average even in the professional identity treatment. For example, if one-third of scientists would lie partially by over-reporting one tail-step, as the anonymous survey evidence cited above might suggest, we would expect an average tail toss reporting of 2.31 tails, leading to

**Hypothesis 3:** Even in the professional identity treatment, average reporting behavior differs from the truthful distribution.

As part of a comprehensive analysis of truth-telling behavior of scientists in the next section, we will confront these hypotheses with our experimental data.
3. Results

We have received 599 responses to the survey, amounting to a response rate of more than 30%.

437 responses contain a coin toss report. Participants come from all major continents, predominantly from Europe and North America. There are 58% male subjects in our sample. The mean age of our subjects is 43 years, and 52% of our participants have a tenured position.

Before we turn to scrutinizing the decisions in the coin-tossing task, we test whether our implicit measure of identity priming using the word completion task indicates that priming has been successful. For each subject, we aggregate over the given numbers assigned to completed words for the six potential word checks (1 for words associated with professional life, 0 for words associated with private life) and compare the mean value of these aggregate numbers for the two treatments. Furthermore, we create an index that captures the relative frequency of mentioning words associated with professional life. We find that the mean number of ‘professional’ words, such as “journal”, “paper” or “session”, is with 2.89 higher in Professional as compared to the 2.66 ‘professional’ words in Private (t-test: p = 0.053). Furthermore, the relative frequency of mentioning words associated with professional life is higher in Professional, with 59%, as compared to Private, with 55% (t-test: p = 0.088). We therefore find some supportive evidence that our Professional treatment was able to make the professional scientific identity of our subjects more salient compared to the Private treatment.

We now examine the coin toss reporting behavior of scientists. Figure 1 shows the theoretical binomial distribution for four tosses of a fair coin (blue dots connected by the dashed line), which is the distribution that we would expect if all subjects report the outcome of their four coin tosses truthfully. The probability that four times tossing a coin results in \( t_i = 0 \) or 4 (1 or 3) [2] times tails is 6.25% (25%) [37.5%]. We refer to this distribution as the ‘truthful distribution’, with a mean truthful response of \( \bar{R}_t = 2 \) tail tosses. The mean payoff-maximizing choice would be the reporting of \( \bar{R}_p = 4 \) tail tosses.

---

20 Overall, 946 individuals clicked on the link to our study. We dropped 10 observations because they responded more than once and one observation because we could identify her as still being a master student. 162 subjects completed some parts of the initial demographic questions, priming questions, or the risk task, but did not complete the coin-tossing task. Appendix B provides a comprehensive investigation of potential response bias and selection effects concerning the balance across treatments. We are confident that our main results indeed capture differences due to varying the salience of professional versus private identity and are not driven by response and selection effects.

21 All p-values reported in this paper are based on two-sided tests.
tosses. The colored bars in Figure 1 show actual reporting behavior of the participating scientists across the two treatments: Private and Professional.

First, we analyze overall coin toss reporting of all scientists by aggregating results from both treatments. We find that overall reporting by scientists differs highly significantly from payoff-maximization. Scientists report on average 2.32 tail tosses, thus indicating substantial lying costs. However, we also find that scientist over-report tail tosses to their advantage: A Kolmogorov–Smirnov test for comparing overall reporting behavior against the binomial distribution confirms that scientists over-report tail tosses ($p < 0.001$). We therefore cannot reject Hypothesis 1 and previous findings in the literature also for scientists.

![Bar Chart](image)

**Figure 1:** Tail toss-reporting behavior of behavior of scientists in the Private identity treatment (red bars) versus the Professional identity treatment (green bars). The blue, dashed line with dots corresponds to the expected distribution if every scientist reported the true outcomes of their coin tosses. The payoff was five Euros times the number of tails reported, i.e. the payoff-maximizing reporting was four times tails.

We now analyze truth-telling in our two treatments. Figure 1 shows reporting behavior of scientists in the private compared to the professional identity treatment. Subjects in Private report 2.41 tail tosses on average, which is higher than the average report of subjects in Professional of 2.24 tail tosses ($t$-test: $p = 0.073$). In particular, we find that scientists in Professional report fewer four times tails as compared to those in Private.
(9.21% vs. 16.16%; chi-squared test: p = 0.028). This confirms our central Hypothesis 2 and establishes

Result 1: Reporting behavior under identity priming

Scientists in the professional identity treatment report, on average, lower tail tosses compared to those in the private identity treatment.

Even though there is fewer over-reporting of higher tail tosses among scientists in Professional compared to the Private control treatment, we still find that there is over-reporting of tail tosses among those primed with their professional identity: A Kolmogorov–Smirnov test for comparing overall reporting behavior in Professional against the expected truthful binomial distribution rejects the null hypothesis at p < 0.01. That is, the coin-toss reporting in Professional still deviates from the truthful distribution, thus confirming Hypothesis 3. Summarizing this finding yields

Result 2: Reporting behavior in the Professional identity treatment compared to the truthful distribution

Scientists in the professional identity treatment over-report tail tosses compared to the expected truthful distribution.

As the marginal behavioral impact of increasing the salience of the professional or private identity will depend on the individual baseline salience level (cf. Benjamin et al., 2010), we make use of having inquired about the participant’s location when completing the survey to explore differences in reporting behavior across locational contexts.22 We compare responses of participants who respondent from their usual workplace “at work” (n=252) with those being “not_at_work”, composed of “at home” as well as “home office” (n=139). We find that the identity priming treatment effect is particularly strong for those scientists responding while not being at their workplace. While the mean number of ‘professional’ words in Private is with 2.65 roughly the same as for the whole sample, we find that the mean number of ‘professional’ words in Professional is 3.11 and thus considerably higher than in Private (t-test: p = 0.044). While there is no tail toss reporting difference across treatments for scientists responding from their workplace (t-test: p = 0.821), the priming intervention had a particularly strong

22 Pre-offered options were “at work”, “at home”, and “home office”, and a residual “other” option.
effect on tail toss reporting for those that were not at their usual workplace (at home, home office, on travel, on vacation etc.): Average tail tosses reported are 2.53 in Private and 2.10 in Professional (t-test: p = 0.008). For four times tails reporting, we find relative frequencies of 18.18 % in Private and 4.11 % in Professional (t-test: p = 0.007).

Result 3: Identity priming and coin toss reporting effects at different locations

The professional identity priming and treatment effect on lower over-reporting is particularly pronounced when participants respond from locations other than their usual workplace. 21 We further relate tail toss reporting to the two other behavioral measures that we collected: risk preferences and donations. 24 First, we elicited risk preferences using the so-called Eckel-Grossmann task (Binswanger, 1981; Eckel and Grossman, 2002). Unlike a number of previous studies that examined the relationship between risk-taking and truth-telling, 25 we find that higher tails reporting is associated with higher risk-taking (correlation-coefficient: -0.13; t-test: p < 0.007). 26 We explore the effects of professional identity priming on risk-taking behavior of scientists in more detail in a companion paper (Drupp et al., 2017). As there is no significant difference in the overall identity priming treatment effect on risk-taking, we are confident that the negative correlation between risk-taking and truth-telling is not driving the key results of the present paper.

Second, we allowed participants to donate fractions (in 10 % steps) of their earnings at the end of the experiment to the NGO ‘Doctors Without Borders’, providing us with an eleven-point step measure of the payoff-fraction donated. This option was not announced earlier, so their donation decision could not have impacted tail toss reporting, but their coin toss reporting and resulting pay-off level might have impacted subsequent donations. We find that participants reporting higher tail tosses are associated with lower step-level donations (correlation-coefficient: -0.17; t-test: p = 0.001). Indeed, the donation fraction decreases monotonically with reported tail tosses (from 94% for 0 tail tosses to 52% for 4 tail tosses). Yet, we find that the absolute donation amount increases

Note that as the variables “at_work” and “treatment” are not significantly correlated (t-test: p > 0.55); this locational effect does not drive our main treatment effect. 21 Tail toss reporting is not associated with participants’ elicited degree of social time preference (t-test: p > 0.70). The same holds for the year of birth (p > 0.70), gender (p > 0.90), being married (p > 0.15), and having tenure (p > 0.35), as revealed by two-sided t-tests. 22 For example, Abeler et al. (2014), who rely on a stated preference measure for the German population, or Drupp et al. (2016), who use the same Eckel-Grossmann risk-elicitation task. 23 Zimerman et al. (2014) examine the relationship between a stated-preference measure of risk-taking specifically in the domain of ethical risks and find that the stated measure of risk-taking in ethical context is positively correlated with dishonest behavior as elicited using a coin tossing task.
monotonically with reported tail tosses (from 11 € for 0 tail tosses to 17 € for 4 tail tosses), resulting from higher pay-offs for people with higher reported tail tosses (t-test: \( p = 0.004 \)). Furthermore, we find that those who do not donate at all report on average 2.50 tail tosses as compared to only 2.17 tail tosses for those who donate all of their pay-off (t-test: \( p = 0.009 \)). Overall, this suggests some consistency of behavior related to social preferences as revealed by both truth-telling and donation levels and yields

**Result 4: Relationship between reporting behavior and donations**

*Lower over-reporting of tail tosses is, on average, associated with a higher share of subsequent donations.*

Finally, we relate tail toss reporting to observable measures of scientific output or scientific success: the number of citations, the h-index and publications as listed in the research database SCOPUS.\(^{28}\) We find that the number of publications, the h-index as well as the number of citations are associated with lower over-reporting of winning tail tosses (t-tests, controlling for age: \( p = 0.079 \), \( p=0.084 \) and \( p = 0.096 \) respectively).\(^{29}\) Furthermore, we have gathered data on whether our participants have published in the prestigious science journals *Science*, *Nature* and *PNAS*. The 36 participants who have published in these journals only report 2.06 tails on average and thus tend to report fewer winning tail tosses (t-test, controlling for age: \( p = 0.106 \)).\(^{30}\) Figure 2 depicts these findings using linear fit. We therefore conclude that honest reporting tends to be related to having a higher scientific output or success. We summarize

**Result 5: Relationship between reporting behavior and scientific output**

*Lower over-reporting of tail tosses is associated with higher measures of scientific output.*

---

\(^{27}\) We find no difference in fractions donated across *Private* and *Professional* (\( p > 0.60 \)). Also for the absolute donation amount we find no differences across treatments (\( p > 0.35 \)).

\(^{28}\) We find no difference in observable measures of scientific output across *Private* and *Professional* for both number of publications (t-test: \( p > 0.65 \)) and citations (t-test: \( p > 0.85 \)).

\(^{29}\) Without conditioning on age, the p-values for the number of publications, h-index and the number of citations are \( p = 0.069 \), \( p = 0.161 \) and \( p = 0.153 \), respectively.

\(^{30}\) Without conditioning on age, the p-value is \( p = 0.100 \). We have made the conservative assumption that for those for whom we could not obtain information on whether she has published in either of these three journals has not published there. If we disregard those for whom we could not obtain this information, the p-value (when controlling for age) would be \( p = 0.076 \) (\( p = 0.079 \)).
Figure 3: Tail toss-reporting behavior and measures of scientific output gathered in the search engine SCOPUS: The number of citations and publications, the h-index as well as having published in the journals Science, Nature and PNAS.

4. Discussion and conclusion

We have investigated whether scientists tell the truth by means of an incentivized coin-toss truth-telling task within an online field experiment with 437 members of an international scientific organization. In particular, we compare truth-telling behavior, in the form of coin toss reporting, across two treatments that either made participants’ professional or private identity more salient using nine identity priming questions. Our key result is that significantly fewer scientists over-report winning tail tosses in the professional identity treatment. This professional priming treatment effect on truth-telling is particularly strong for those responding to the study while not being at their usual workplace. Furthermore, we show that truth-telling is positively associated with measures of a scientist’s output, such as the number of citations and publications as well as having published in the leading science journals Nature, Science and PNAS. We therefore conclude that honest reporting tends to be related to having scientific success. While this is no causal evidence, our findings do not support claims that misbehavior is more prevalent among successful scientists and one may even conjecture that science rewards honest behavior.

While we are able to provide causal evidence that professional identity effects associated with science foster truth-telling, we can pinpoint the underlying mechanism
for this finding only inductively. 31 Previous work that our simple model of truth-telling behavior builds upon (Benjamin et al., 2010; Cohn et al., 2015) suggests that this more frequent truth-telling is driven by stronger honesty norms associated with the professional (in this case scientific) identity. This main interpretation would suggest that academia fosters a culture of truth-telling that is consistent with its general aim of searching for truths. Indeed, this cultural norm-based interpretation has featured prominently in related findings in experimental studies on the banking industry (Cohn et al., 2014; Vileval, 2014). Yet, stronger honesty norms may not be the only facet of the professional identity of scientists that drives truth-telling behavior. For example, it is often suggested that competitiveness (‘publish or perish’) is a central feature of behavioral patterns and thus perhaps also associated norms in academia (see, e.g., Fanelli, 2010; Necker, 2014). If this were the case, our main treatment effect finding would be a conservative estimate of the truth-telling norms that science nurtures, as also inherent competitiveness norms might have a detrimental effect on truth-telling. 32

Besides the interpretation that honesty norms associated with the scientific identity drive truth-telling behavior, it could also be the case that other professional identity concerns may impact our results. Specifically, it could be that scientists strategically report more honestly as they might seek to paint a more positive picture of science. That is, they may take reputational concerns at the level of the profession into account. 33 We regard this alternative explanation as an unlikely mechanism. A necessary condition for this strategic influence explanation is that participating scientists believe that they can favorably influence the overall outcome, i.e. their contribution is non-marginal. The participants in our study knew that we targeted all members of the scientific organization, i.e. 1/n was small. Given our between-subjects design, participants were also not aware that there was another treatment. 34 Thus, even though

31 Taking the study by Cohn et al. (2014) as an example, Vranka and Houdek (2015) discuss the difficulty of pinpointing underlying mechanisms of observed priming effects.
32 For example, Shleifer (2004) discusses how (market) competition may have detrimental effects on ethical behavior. More recently, a series of experimental economic studies have found that competition may lead to more dishonesty (see, e.g., Cartwright and Menezes, 2014; Conrads et al., 2014; Faravelli et al., 2015; Rigdon and D’Estere, 2015; Schwieren and Weichselbaumer, 2010). However, while Fanelli et al. (2015) find that scientific misconduct is more likely in countries where individual research output yields monetary rewards, their results do not support the hypothesis that pressure to publish seems to drive dishonest behavior. Furthermore, Cohn et al. (2014) do not find an identity priming effect for bankers on a stated preference question on competitiveness.
33 This strategic behavior could thus be present in both treatments, but due to our experimentally induced higher salience it would likely be higher in the professional identity treatment.
34 Furthermore, while truth-telling approaches are well-known in behavioral economics and psychology by now, the participating natural scientists had very limited exposure to such experiments.
we cannot rule out the presence of professional reputation concerns by design, it seems rather unlikely that this will be a main driver of our observed treatment effect.  

While our central treatment effect therefore seems to suggest that science fosters a culture of honesty, which is arguably good news for science as well as for all of us relying on scientific results, we still find that scientists over-report winning tail tosses compared to the expected truthful distribution. For the whole sample, we find that scientists report on average 2.32 tail tosses. In the professional identity treatment, we also find that scientists over-report winning tail tosses. Thus, even though we find that scientists who report tail tosses more honestly on average tend to be more successful scientifically, the culture of honesty that academia is built on does not seem sufficient to ensure that science does not get derailed from its quest for truths. This finding is in line with the anonymous survey based approaches that provide evidence that a considerable fraction of scientists engage in questionable research practices (see, e.g., Fanelli, 2009; John et al., 2012; List et al., 2001; Martinson et al., 2005; Necker, 2014).  

As scientific honesty is crucial for scientific development as well as the public’s trust in the results of science, further measures have to be taken to prevent scientific misconduct. Meta-analyses (Abeler et al., 2016; Brodeur et al., 2016), replication studies (Camerer et al., 2016; Dreber et al., 2015; Open Science Collaboration, 2015), more precise and transparent reporting practices (Christensen and Miguel, 2016; Miguel et al., 2014; Nosek et al., 2015; Simmons et al., 2011) as well as institutional incentives and arrangement for research integrity (Titus et al., 2008; Titus and Bosch, 2010) are some important recent steps into this direction. Besides showing that academia already seems to foster norms associated with truth-telling, our findings thus call for further steps that let this quest for improving research conditions and practices continue.

---

35 Furthermore, if portraying a positive image of science would drive our treatment effect in truth-telling behavior, one might also expect that such strategic behavior to show up in subsequent donation decisions. Yet, we find not significant differences across the two treatments for both the fraction of pay-off reported and for the absolute size of donations.

36 For example, this result would occur if one-third of scientists lied partially, as is suggested in a number of anonymous surveys on scientific misbehavior.
References


Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS One, 10*(6), e0127556.


Appendix A: Screenshots from the online survey

Figure A.1: Priming questions for the *Private* identity treatment

![Image of priming questions](image_url)
Figure A.2: Priming questions for the Professional identity treatment
Appendix B: Testing for response and selection bias

Laboratory experiments implicitly constrain participants to make choices and remain in the laboratory for the entire length of a study in order to complete it. Conversely, (online) field experiments potentially suffer from response bias and attrition.

To test for obvious response bias, we carry out several checks suggested in the previous literature. In particular, we test whether there are observable differences for early versus late respondents (e.g. Necker 2014) as well as consider observable characteristics of our respondents and non-respondents. First, for earlier versus later respondents, we do not find significant differences in tail toss reporting between the first half, with a mean tail toss of 2.33, and the second half of respondents, with a mean tail toss of 2.31 (t-test: p = 0.847).

Second, we compare observable characteristics of our 437 respondents who completed the coin-tossing task and those who dropped out of the study that we still have some information on (see Table B.1).1 There are no significant differences across participants and dropouts except for their age: Those participating in the coin toss experiment are 4.53 years younger than those dropping out (t-test: p = 0.000). As age is not significantly correlated with overall reporting behavior among participants (t-test: p = 0.747), this does not provide an indication for obvious response bias.

---

1 On those who have dropped out, we have information on the assigned treatment as well as the continent on which they were located when clicking on the participation link for 244 drop-outs, and on gender and their mean year born for 162, and their experimental risk choices for 39 drop-outs.
Table B.1: Descriptive statistics for participants and drop-outs

<table>
<thead>
<tr>
<th></th>
<th>Coin toss</th>
<th>No coin toss</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 437</td>
<td>n = 244/162/39</td>
<td></td>
</tr>
<tr>
<td>Share from Europe</td>
<td>0.78</td>
<td>0.81</td>
<td>0.377</td>
</tr>
<tr>
<td>Share Professional</td>
<td>0.55</td>
<td>0.52</td>
<td>0.507</td>
</tr>
<tr>
<td>Mean year born</td>
<td>1972.85</td>
<td>1968.32</td>
<td>0.000</td>
</tr>
<tr>
<td>Share male</td>
<td>0.59</td>
<td>0.63</td>
<td>0.311</td>
</tr>
<tr>
<td>Mean risk choice</td>
<td>3.92</td>
<td>3.46</td>
<td>0.135</td>
</tr>
</tbody>
</table>

Note: The p-values for binary data are based on chi-squared tests and the p-values for interval data are based on t-tests.

We further examine balance across our experimental treatments. For this, we compare *Professional* and *Private* for observable information that we collected in both treatments. We know that the computer-generated randomization roughly worked: about one half, 52.85%, of the 946 clicks on the e-mail’s invitation link were randomly assigned to *Professional* and the remainder to *Private*. Compared to the 52.85% who were assigned to *Professional* when they clicked the invitation link in the e-mail, we have 54.69% (239 out of 437) of participants who remained in *Professional* and completed the coin-tossing task. The numbers point to slightly greater attrition in *Private* compared to *Professional*. Table B.2 shows further descriptive statistics for the participants who completed all subsequent stages of our study including the coin-tossing task.

Table B.2: Descriptive statistics.

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th><em>Professional</em> treatment</th>
<th><em>Private</em> treatment</th>
<th>p-values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 437</td>
<td>n = 239</td>
<td>n = 198</td>
<td></td>
</tr>
<tr>
<td>Share from Europe</td>
<td>0.78</td>
<td>0.80</td>
<td>0.76</td>
<td>0.416</td>
</tr>
<tr>
<td>Mean year born</td>
<td>1972.85</td>
<td>1972.25</td>
<td>1973.58</td>
<td>0.219 / 0.180</td>
</tr>
<tr>
<td>Share male</td>
<td>0.59</td>
<td>0.54</td>
<td>0.64</td>
<td>0.032</td>
</tr>
<tr>
<td>Share tenured</td>
<td>0.52</td>
<td>0.50</td>
<td>0.53</td>
<td>0.544</td>
</tr>
<tr>
<td>Share “at work“</td>
<td>0.65</td>
<td>0.66</td>
<td>0.63</td>
<td>0.564</td>
</tr>
<tr>
<td>Mean risk choice</td>
<td>3.92</td>
<td>3.89</td>
<td>3.96</td>
<td>0.677 / 0.656</td>
</tr>
</tbody>
</table>

Note: The p-values for binary data are based on chi-squared tests and the p-values for interval data are based on two-sided t-tests / rank-sum tests.
On average, the participants in the study were born in 1973, meaning that—as of 2016—they were 43 years old on average. Around half of the participants held tenured positions. 20% lived in the US, while 78% lived in Europe. 59% of the participants are male, the rest is female. Comparing the characteristics across treatments shows that our treatments are balanced, except for gender. The share of males in Professional is 54% compared to 64% in Private (chi-squared test: $p = 0.032$). As we find that gender is not significantly correlated with overall tail toss reporting behavior (chi-squared test: $p = 0.588$), this does not appear as problematic at first sight, especially given that in our between-subjects design it was not possible for subjects to actively select themselves into any treatment. Further, they did not know that a second treatment existed. However, the main treatment effect in Result 1 is particularly pronounced for males: We find that there are no significant differences in overall reporting behavior across the 254 male and 181 female subjects: mean tail toss reports are 2.32 and 2.30 tails respectively (chi-squared test: $p = 0.588$). However, there are differences in the treatment effect across gender: While there is no significant difference in reporting behavior of females across the identity priming treatments (t-test: $p = 0.695$), male participants significantly over-report tail tosses in Private compared to Professional (t-test: $p = 0.061$). It is therefore worthwhile to explore potential explanations of this gender balance difference in more detail.

Fortunately, we have information on the gender distribution in our population (the e-mail list of the scientific organization). We know that about 66% of the members in the population are male. This figure is very close to the 64% of males in Private (binominal probability test, $p = 0.497$). Thus, there are significantly fewer males in Professional compared to the expected 66% (binomial probability test, $p < 0.001$). In other words, we find the expected share of males in the Private treatment, while there are significantly fewer males and conversely relatively more females in Professional than expected. We do not have detailed information on most dropouts, as these occurred before subjects provided any information in the survey. However, we can extend the analysis of dropouts above to consider differences across treatments within the dropouts.

First, the sequential nature of our experimental tasks allows comparing the risk-taking behavior of those 39 participants who have completed the risk elicitation task but not the coin-tossing task. Among these 15 are from the Private and 24 from the Professional treatment, i.e. we had somewhat greater attrition in Professional. Those in Private not completing the coin-tossing task had a mean risk choice of 4.00. Those in Professional.

\footnote{Furthermore, chi-squared tests: $p > 0.40$ for all single tail tosses.}
had a mean risk choice of 3.13. Although this difference is not significant due to the small number of observations (t-test: p = 0.210), as higher risk choices are significantly correlated with lower truth-telling, if at all this may suggest that our observed main treatment effect may be a conservative estimate.

Next, we consider dropout rates across gender per treatment (see Table B.3). For this, we consider all dropouts for whom we have information on their gender and divide this by the respective combined number of dropouts and tail toss respondents. We find that overall and also across both genders there are higher dropout rates in Private as compared to Professional. Furthermore, we find more frequent attrition of males, as compared to females, yet this difference is not significant (see Table B.1).

### Table B.3: Dropout rates per treatment and gender

<table>
<thead>
<tr>
<th>Gender</th>
<th>Private treatment</th>
<th>Professional treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>0.30</td>
<td>0.27</td>
</tr>
<tr>
<td>Female</td>
<td>0.26</td>
<td>0.24</td>
</tr>
<tr>
<td>Overall</td>
<td>0.29</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Note: 54 (25) males (females) in Private and 48 (35) males (females) in Professional dropped out of the study.

This analysis of dropouts therefore cannot explain why we find significantly fewer males and conversely more females in Professional than expected. It thus seems that the more frequent relative participation of females in the Professional treatment occurs at a stage that precedes our experimental treatments and thus cannot be driven by a selection effect of females or males into the treatments. We refrain from speculating about these males’ reasons for dropping out or those females participating more frequently.

What we can do, however, is to explore the robustness of our results by means of simulations. Table B.2 has shown that there are only 54 % males in Professional, as compared to 64 % in Private. For our simulations, we therefore hypothetically add another 25 males to the Professional treatment, such that the proportion of males would be equalized across treatments to 64 %. We consider five cases that assume different distributions of lying behavior for those 25 additional males. They would report: First, as males in the Professional treatment (Simulation 1); Second, as all of those in the Professional treatment (Simulation 2); Third, as all respondents across both treatments (Simulation 3); Fourth, as all those in the Private treatment (Simulation 4); Finally, they would report on average as the group with the highest overall lying behavior: males in the Private treatment (Simulation 5). These different simulations (summarized in Table B.4) thus add
observations whose tail toss reporting is shifted to the right by varying degrees as compared to the expected truthful distribution.3

Table B.4: Treatment differences (p-values) in tail toss reporting across Private and Professional for our respondents and three simulations with additional males

<table>
<thead>
<tr>
<th></th>
<th>Overall tail tosses</th>
<th>4 times tail tosses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p-values</td>
<td>p-values</td>
</tr>
<tr>
<td>Original participants</td>
<td>0.073</td>
<td>0.028</td>
</tr>
<tr>
<td>Simulation 1</td>
<td>0.067</td>
<td>0.021</td>
</tr>
<tr>
<td>Simulation 2</td>
<td>0.061</td>
<td>0.021</td>
</tr>
<tr>
<td>Simulation 3</td>
<td>0.074</td>
<td>0.030</td>
</tr>
<tr>
<td>Simulation 4</td>
<td>0.089</td>
<td>0.043</td>
</tr>
<tr>
<td>Simulation 5</td>
<td>0.098</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Note: The p-values for overall tail tosses are based on t-tests and the p-values for the 4 times tail tosses are based on chi-squared tests.

We find that the treatment effect in terms of overall truth-telling behavior is qualitatively robust across all simulations when considering a t-test (p < 0.10).4 For the difference in four tails reporting we find that the treatment effect is qualitatively robust across Simulations 1-4 (p < 0.05). For Simulation 5 we still find a significant treatment effect at p = 0.059, i.e. at p < 0.10.

Overall, this simulation exercise suggests that those 25 `statistically missing' males in the Professional treatment would have to be substantially less honest as our respondents such that selection would drive our treatment effect. Thus, although it is not possible to rule out selection and response bias in field experiments due to attrition, we are confident that our main results indeed capture differences due to varying the salience of professional versus private identity and are not driven by response and selection effects.

3 The number of 0/1/2/3/4 tail tosses for these three cases are as follows: 1/4/10/8/2 for as males in Professional, 1/4/11/7/2 for as in Professional, 1/4/10/7/3 for as in overall, 1/4/9/7/4 for as in Private, and 1/4/9/6/5 for as males in Private (this compares to 2/6/9/6/2 in the expected truthful distribution).

4 Note however that when considering a rank-sum test, the treatment effect is not robust for Simulations 4 and 5 (with p = 0.109 and p = 0.116, respectively).