Self-Serving Truthfulness *

Stephanie A. Heger†
School of Economics, The University of East Anglia

Robert Slonim
School of Economics, The University of Sydney

Franziska Tausch

July 23, 2020

Abstract: We conjecture and experimentally examine a novel explanation for dishonesty, *self-serving truthfulness*. We conjecture that when the objective truth is uncertain, people will provide self-reports that financially benefit them. To examine *self-serving truthfulness*, we asked a sample of U.S. car owners to respond to an auto insurance underwriting questionnaire that affects their price of insurance (i.e., premium), and investigated how financial incentives affect the honesty of their responses. We find, consistent with the current literature, that people have a strong preference for truthfulness, but only when they are confident of the objective truth. However, when people are not completely certain of the objectively correct response, significant dishonesty occurs in a self-serving manner. We also find that self-serving truthfulness is not strategic and thus distinct from self-deception. Further, interventions aimed at mitigating dishonesty by targeting commonly used strategies to justify dishonesty, e.g., moral wiggle room, consequence attenuation, and attenuation through victim deservingness, are minimally effective.

*This project benefitted from many helpful comments and conversations.
†[Corresponding author] Email:saheger@gmail.com

1 Introduction

Dishonest, fraudulent and corrupt behaviors pose substantial costs to many private and public organizations and societal wellbeing—international costs of corruption total $3.6 trillion or more than 5% of global GDP (World Economic Forum, 2018) and 1 in 4 adults across the world report having to pay a bribe to access public services in the past year (International Transparency, 2017). Becker (1968) argued that the decision to commit crime (including fraud and corruption) may be based on a rational cost-benefit analysis. However, while the standard economic model predicts individuals to behave as cold cost-benefit calculators, more recent literature suggests that individual's willingness to cheat, lie and commit fraud for financial gain depends on whether they are perceived by others, and by themselves, as dishonest (Abeler, Nosenzo, and Raymond, 2019; Cressey, 1986; Mazar, Amir, and Ariely, 2008). This suggests that people weigh the benefits from unethical behaviour against the costs, which also include the desire to maintain a moral self-image (Rabin, 1995; Ayal et al., 2015)—a tendency that emerges in childhood (Maggian and Villeval, 2016).

To maintain their positive image while behaving dishonestly, people frequently use behavioural strategies to justify, excuse or remain ignorant about the consequences of their dishonest behaviour (Sykes and Matza, 1957; Ayal et al., 2015; Bandura, 1999; Shu, Gino, and Bazerman, 2011; Shalvi et al., 2015; Köneke et al., 2015), resulting in a reduction of "ethical dissonance" (Ayal et al., 2015). In particular, there is a growing evidence that individuals use uncertainty, ambiguity and subjectivity when behaving dishonestly to preserve their self image (Kunda, 1990; Konow, 2000; Dana, Weber, and Kuang, 2007; Mazar, Amir, and Ariely, 2008; Haisley and Weber, 2010; Shalvi et al., 2015; Exley, 2016; Grossman and Van Der Weele, 2017; Gneezy et al., 2020). For example, ambiguity regarding the appropriate moral behavior may play a central role in justifying dishonest behavior as individuals can use this ambiguity to mask misbehavior behind a veil of ignorance (Shalvi et al., 2015; Mazar, Amir, and Ariely, 2008). Gneezy et al. (2020) focus on the role of self-deception, or the strategic manipulation of beliefs, to neutralize the effect of dishonest behaviour on one's self-image.

In this paper, we propose an additional and novel explanation in the literature on how the moral costs of dishonesty are neutralized, *self-serving truthfulness*, in which individuals are uncertain of the exact truth of a piece of objective information they are asked to provide and when tasked with making an approximation of the truth, they "fill-in" this uncertainty in a self-serving way.[1] For example, suppose we ask two groups of subjects to report the month of their mother's birthday,

[1]Our notion is related to the literature on elastic justification. Schweitzer and Hsee (2002) report results from an experiment in which subjects were asked to play the role of a hypothetical car salesman. In the experiment, the uncertainty about mileage of the hypothetical car was randomly varied and sellers were more likely to report lower mileage in the high uncertainty condition than in the low uncertainty condition. However, the experiment in Schweitzer and Hsee (2002) was unincentivized and thus it is unclear whether subjects had a preference over their responses.

the month of their mother's mother's birthday and the month of their grandmother's mother's birthday. Subjects are likely to be quite certain of the month of their mother's birthday, less certain of their grandmother's birthday and even less certain of their great grandmother's birthday. One group is given a fixed rate for the answers and the other group earns a bonus payment for reports of birthdays that fall in "even" months. Self-serving truthfulness predicts that as uncertainty about the truth increases, subjects who have an incentive to lie (i.e., those in the second group) will be more likely to report even months. Thus, we expect the probability of reporting an even month for mothers' birthdays to be relatively the same between the groups, but self-serving truthfulness predicts a higher percentage of even birthdays reported for grandmother's and even larger percentage of even birthdays reported for the great-grandmother in the incentivized group relative to the non-incentivized group.

In our experiment, we ask car owners in the U.S. to participate in an online survey that resembles an auto insurance underwriting application. The survey is an 11-item questionnaire that asks participants about driving habits, relevant demographics (see survey here) and later elicits their confidence in the accuracy of their answers to each of the 11 items in the questionnaire. Our study includes three (sets of) conditions. In our Control Treatment, participants are asked to respond with no financial consequences tied to their answers. In our Incentive Treatments, we repeat the same questions, but include high stakes financial incentives that proxy for the cost structure behind premiums in underwriting policies, using two different stake levels. Our third set of conditions, the Intervention Treatments, also include financial incentives but at the same time apply a series of interventions that are meant to address common explanations found in the literature for dishonesty: moral wriggle room, attenuation, and rational lying detection.

To identify the effects of self-serving truthfulness on dishonest reporting, among the 11 questions, we included questions in which the respondent will definitely know the answer, such as their age and gender, as well as questions we anticipated that the respondent would be much less certain of the correct answer, such as the current value of their car. In order to know which questions subjects were not certain of the correct response, after they completed the underwriting survey, we asked them to indicate the confidence in the correctness of each of their responses. We hypothesize that if participants are subject to self-serving truthfulness, then we will observe more dishonesty on the responses that respondents were less certain about when there are financial incentives compared to when there are not.[2]

The experimental design allows us to answer three main questions: (1) Do incentives increase dishonesty? (2) Does uncertainty about the objective truthfulness of their response increase dishonesty (i.e., is there self-serving truthfulness)? and (3) Can self-serving truthfulness be "de-biased" using standard behavioral nudges intended to address alternative justifications for

[2]This method for identifying dishonesty or truthfulness is similar to John, Loewenstein, and Prelec (2012), thus while we cannot know the truthfulness of any individual response, we can identify how responses change when there are financial incentives.

dishonesty?

We thus present three main findings. First, we find that relative to our Control Treatment, subjects in the Incentive Treatments are significantly more dishonest. On average, when participants have an incentive to lie, they add approximately 30 USD to their payoff in our Base Incentive Treatment versus what subjects in our Control Treatment would have earned had they also been paid bonuses for their responses. The incentives in the Base Incentive Treatment were structured such that subjects could add or subtract \$10 for each change in response (see survey here.) and thus a \$30 increase is equivalent to three minimal lies. [3]

Second, we find evidence consistent with self-serving truthfulness; that is, the majority of dishonesty detected between the Control Treatment and the Base Incentive Treatment is driven by the two questions in which subjects report significantly less confidence in the accuracy of their answers.

Specifically, we find that there is significantly more dishonesty for the each of the two questions subjects were not confident of the objective truth compared to every other question in which they were confident of their responses. For the two questions that the subjects were less confident about, subjects, on average, made a minimal lie (i.e., increased their bonus by \$10), whereas for the other nine questions in which subjects were confident, they lied on average by approximately 1/10 of a minimal lie. In other words, lying appears to be approximately 10 times larger when subjects are less confident in their responses than when they are almost certain of the correct response.

We further find that the beliefs about accuracy reported in the Base Incentive Treatment and the Control Treatment are not statistically different. Thus, we find no evidence that self-serving truthfulness is *strategic* or a form of self-deception (Gneezy et al., 2020). In other words, individuals do not manipulate their reported beliefs when there is an incentive to lie. Instead, self-serving truthfulness appears to be nonstrategic.

To address our third research question, we distinguish self-serving truthfulness from other popular strategies used to justify dishonesty: moral wiggle room, rational lying and consequence attenuation. If dishonest behavior is driven by these three strategies, then we anticipate that we will see less lying with our interventions. On the other hand, if self-serving truthfulness is driving dishonesty, then we hypothesize that these strategies will not mitigate dishonest behavior.[4]

Like self-serving truthfulness, moral wiggle room strategies use uncertainty to justify questionable or dishonest behavior. However, self-serving truthfulness stems from an uncertainty about the objective truth, whereas moral wiggle room applies in scenarios where there is uncertainty about

[3]Since we cannot measure whether subjects are telling the truth in the Control Treatment, the Base Incentive Treatment only captures the additional dishonesty that stems from financial incentives. In the Control Treatment, individuals may also be motivated to lie due to social desirability and image concerns.

[4]See (Köneke et al., 2015) who provide a comprehensive overview of justification strategies applied in the insurance context.

what types of behaviors are morally acceptable. To intervene on dishonesty stemming from moral wiggle room, we design a set of treatments that explicitly informs subjects about the types of behaviors that are deemed unacceptable and dishonest. Further, we ask subjects to agree to abide by an honor code of honesty. We find, consistent with our interpretation of self-serving truthfulness as distinct from moral wiggle room, that our Moral Wiggle Room interventions are only minimally effective at mitigating the dishonesty we detect in the Base Incentive Treatment. More specifically, the three moral wiggle room interventions pooled together decrease the dishonesty detected in the Base Incentive Treatment by only 24%.

We also examine rational lying and consequence attenuation strategies. A rational lying strategy occurs when an individual weighs the costs and the benefits of behaving dishonestly and does so when the benefits outweigh the cost (Becker, 1968). In the Rational Lying Treatment, subjects are reminded that there can be costs if an individual is caught lying. A consequence attenuation strategy occurs when people find ways of using extenuating circumstances to justify dishonest acts, such as denying that their misbehavior has any great, harmful consequences or that a true victim exists (Sykes and Matza, 1957; Köneke et al., 2015). The latter may involve strategies such as dismissing the victim as simply being a 'faceless' corporation or institution that can afford to be harmed, evaluating the victim as deserving harm based on its previous bad deeds or exhibited immoral behaviours, or pointing to perceived inequalities or unfairness that dishonest behaviour might 'correct' (Shalvi et al., 2015; Bellé and Cantarelli, 2017; Fukukawa, 2002). We experimentally examine consequence attenuation strategies using two treatments: (1) the Victim Treatment reminds subjects that other individuals are hurt by their dishonesty and (2) the Image Treatment highlights that the insurer has been engaged in a variety of socially beneficial deeds.[5] We find that the Rational Lying Treatment is ineffective at mitigating dishonesty, while the Consequence Attenuation interventions reduce dishonesty by only 9%.[6] Thus, overall our three intervention treatments to address common explanations for dishonesty, moral wiggle room, rational lying and attenuation, only minimally reduce dishonestly, which is consistent with our alternative explanation for lying occurring due to self-serving truthfulness.

2 Experimental Design

We recruited 1,530 participants via the online platform Amazon Mechanical Turk (Mturk) to participate in a survey that resembles an insurance underwriting questionnaire. All participants

[5]See also Christodoulou et al (2020) who test the effectiveness of consequence attenuation and attenuation through victim deservingness in a field experiment with an insurer.

[6]While we discuss the results from the Rational Lying and Consequence Attenuation interventions, we relegate the results to Appendix B for two reasons because: (1) these types of interventions are not as developed in the literature and thus there may be specific interventions aimed at these two strategies that are not included here but that might yield different results; and (2) these interventions are more hypothetical and thus we interpret the results cautiously.

are car owners and located in the US. They all earned 1 US Dollar for finishing the survey, and five participants are randomly drawn to be paid an additional bonus. This bonus starts at $350 and – depending on the Treatment – may be adjusted depending on the responses the participant indicates in the questionnaire. The bonus payments were $354 on average, ranging from $80 to $500. All Treatments were released on Mturk at the same time. The experimental instructions used for the two main treatments, the Control Treatment and the Base Incentive Treatment can be found here.

The questionnaire includes questions that are typically asked in car insurance underwriting that are used to determine the premium a person has to pay to receive insurance. Table 1 displays a summary of the questions that are asked and the appendix includes the exact wording and response options.

Table 1: 11-item Survey Questions

1.	**Speeding** Number of speeding fines you received in the last five years
2.	**Parking** Your parking habits on and off the street
3.	**Other** Frequency of other drivers with less than five years' experience driving your car in the last year
4.	**Accidents** Number of accidents or incidents involving loss or damage in the last ten years
5.	**Alcohol** Average number of alcoholic drinks you consume in a week
6.	**Miles** Amount of miles you have driven any car in the last five years
7.	**Value** The current value of your car
8.	**Gender** Your gender
9.	**Marital** Your marital status
10.	**Age** Your age
11.	**Licensed** Number of years you have been licensed to drive

2.1 Treatments

For each Treatment condition, the 11 survey questions indicated in Table 1 were presented on a single page of the online survey and always in the same order. Respondents could enter them from top to bottom but could have gone back and forth before answering them all. All 11 questions were forced responses in order to simulate an actual underwriting process in which they would need to answer all questions to be offered a policy.

We implemented three sets of conditions: (1) Control Treatment; (2) Incentive Treatments (Base Incentive Treatment and High Incentive Treatment); and (3) Intervention Treatments. Our Control Treatment provides a baseline for responses to the underwriting questions when there is no monetary incentive to be dishonest. The Incentive Treatments introduce monetary incentives for responses. The Intervention Treatments adds interventions to the Base Incentive Treatment aimed at mitigating dishonesty by weakening the popular justification strategies of moral wiggle room and attenuation. Table 2 lists all of the treatments and sample sizes. Note that since our primary research questions involve comparing the Base Incentive Treatment to every other treatment, we

included twice as many observations in the Base Incentive Treatment to increase power (see List, Sadoff, and Wagner (2011) for a discussion of power for this situation). We describe each set of Treatments below.

Table 2: Sample Sizes by Treatment Assignment

Treatment	Sample Size
Control Treatment	151
Incentive Treatments	
Base Incentive Treatment	308
High Incentive	151
Intervention Treatments	
Moral Wiggle Room	
Signature	154
Signature PS	153
Check Box	152
Attenuation	
Victim	154
Image	153
Rational Lying	152
Total	1,528

Control Treatment In only the *Control Treatment*, participants could receive a fixed potential bonus of $350 which is unaffected by their responses. The insurance context is not mentioned at any point, and participants have no monetary incentive to provide a dishonest response to any question.

Incentive Treatments Participants in the *Base Incentive Treatment* and all subsequent conditions initially receive information that the questions that are going to follow are typically asked in determining the insurance premium that drivers have to pay to receive auto insurance. They are explained how premiums are determined based on whether people are high risk or low risk based on the answers they provide in the questionnaire, and how this in turn affects premiums, and - in a similar manner - the bonus in the experiment. For each question they are provided with information regarding how each response will affect their potential bonus. For 8 of the 11 of our underwriting survey questions, the marginal impact of riskier (from the insurer's perspective) responses decreased the subject's earnings by an additional $10. The remaining 3 of the 11 questions were framed as good driver discounts, where the marginal impact of a less riskier (from the insurer's perspective) responses increased the subject's earnings by an additional $10. Starting with a bonus of $350, the most extreme responses would result in bonus payments that could range from $500 to $0. We also include one variation of the Base Incentive Treatment, called the *High Incentive Treatment*, in which the basic structure is identical, but we triple the amount added to the bonus for the question on people's parking habits.

Intervention Treatments The Intervention Treatments keep the identical structure as the Base Incentive Treatment, but with the additional information described below. The first set of conditions are aimed at dishonesty that stems from moral wiggle room.

Moral Wriggle Room In the *Signature at the Top Treatment* (henceforth: Signature Treatment) participants are asked to confirm an honor statement regarding the truthfulness of their responses by typing their first name. The honor statement is intended to provide a moral cue. This counteracts individuals' strategy to justify dishonesty by referring to a lack of awareness about what behavior is expected from them thereby mitigating 'moral wiggle room'. Signing one's name underneath an honor code has come to be known as an expression of 'social self-presence' - manifesting one's identity on a page as a promise. Shu et al. (2012) find that those who sign such a statement at the top of a document, rather than at the bottom, were more likely to act honestly. E-signatures, however, are found to be less effective. Individuals who gave a handwritten signature indicate they are less likely to breach a contract and are less likely to cheat on simple tasks than those who sign by e-signature (such as PINs, check boxes, or typed names) (Chou, 2015a,b). The evidence on the effectiveness of different types of e-signatures is however limited and mixed, and thus requires further investigation.

In addition to the Signature Treatment, we implemented two additional variations: (1) Signature Previous Screen (henceforth: PS) and (2) Check Box Treatment. In the *Signature PS Treatment*, the honor statement is not on the same page as the questionnaire but on the previous screen. By placing the honor statement on the previous page, participants are required to sign the statement before they can answer the questionnaire. In the *Check Box Treatment*, participants are asked to confirm an honor statement regarding the truthfulness of their responses by clicking a button. This treatment is identical to the Signature Treatment except that subjects only have to check a box rather than type in their name.

Rational Lying A second type of intervention was implemented in the *Detection Treatment*. It is aimed at rational lying strategies and follows the identical structure as the Base Incentive Treatment, except that before the questionnaire we provide the following information to increase the participants' awareness of the risk of their dishonesty being detected in real life: "Once a customer is unfortunate to be involved in an accident and makes a claim, the insurer may verify the information from some of the questions before the claim is validated. If the information turns out to be wrong the customer may lose the entitlement to receive financial help from the insurer." The Detection Treatment is intended to examine Becker's rational detection-punishment approach.

Consequence Attenuation The third type of Intervention Treatments, the Victim Treatment and the Image Treatment, are aimed at counteracting strategies that attenuate the severity

of negative externalities that lying has on others. The structure of these Treatments is identical to the Base Incentive Treatment, except that each of the two Treatments provides additional information before subjects respond to the survey. In the *Victim Treatment* we provide the following information to make participants aware that being dishonest in insurance applications has negative externalities on other policy holders: "Providing an insurer with wrong information can negatively affect other customers. Insurance premiums are calculated based on the personal information that customers provide. If the insurer needs to spend more money on paying claims than expected due to falsely indicated information, the premiums for all customers may be raised." The victim information is intended to make salient that other people are in fact harmed by dishonest reporting. To the extent that dishonest reporting is the result of assuming (or ignoring) the negative externality that others are harmed, making the victims more salient should counterbalance this justification strategy for dishonesty.

In the *Image Treatment*, we provide some (true) examples for socially responsible behavior among automobile insurance companies: "The service of insurance providers often goes beyond providing a security net for its customers. Liberty Mutual, for example, funds non-profit organizations that help the homeless, people with disabilities or those that are in a disadvantaged position, thus supporting the most vulnerable members of society. Likewise, Nationwide awards grants to support organizations that help in emergencies, providing basic needs and crisis stabilization." The image information presented is intended to counteract the strategy that customers may use to justify dishonesty by assuming that insurers are themselves unethical organizations and thus harming them is not immoral. By indicating the good deeds insurers are doing, the image of insurers as immoral is intended to be harder for customers to maintain.

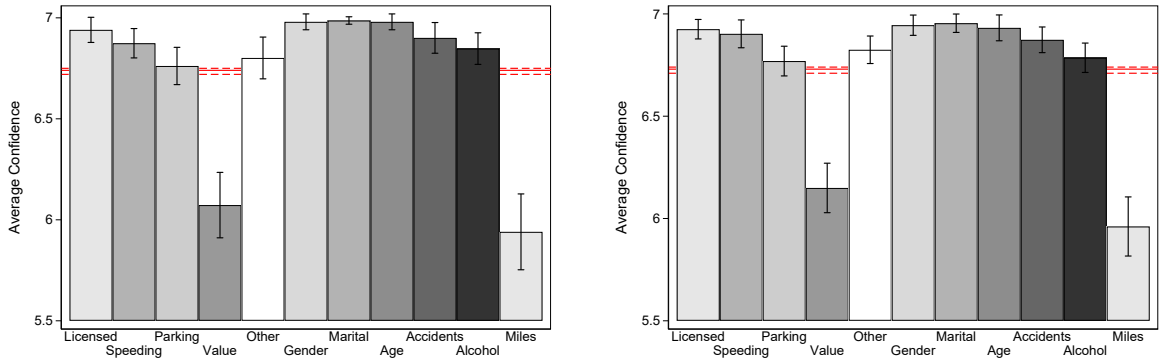
2.2 Measuring Confidence

After subjects completed the 11-item survey we additionally asked them how confident they are in the accuracy of the responses they provided in the survey for each of the 11 questions (the appendix includes the exact wording). When subjects reached the page with the questions on confidence, they were not permitted to return to the previous page to change their responses to the other questions. They were asked to rate their confidence on a scale from 1 (=Not At All Confident) to 7 (=Completely Confident). We conjectured, based on the notion of self-serving truthfulness, that confidence about the correctness of their response would influence honesty.

The average level of confidence reported across the 11 questions was 6.74 (standard deviation= 0.27) in the Control Treatment and 6.73 (standard deviation = 0.46) in the Base Incentive Treatment. We find that subjects are significantly less certain of the accuracy of their answers to the "Miles" and "Value" questions relative to the other 9 questions (mean=6.10 versus mean=6.90, respectively). Figure 1 shows the average level of confidence reported for each of the 11 questions

with 95% confidence intervals. The solid red line represents the average level of confidence across the 11 questions and the dashed red lines represent the 95% confidence interval of the average. As Figure 1 makes clear, not only do subjects report significantly lower levels of confidence in their reports for the "Values" and "Miles" question, but these two questions are the only questions with significantly lower levels of confidence than the average level of confidence. Unpaired t-tests between the confidence reported in the "Values" and "Miles" questions for the Control Treatment and the Base Incentive Treatment indicate significantly lower levels of confidence than the average level of confidence in the two treatments, t=-7.81***, t=-8.16***, t=-8.68***, t=-9.86***, respectively.

FIGURE 1: CONFIDENCE LEVELS



(A) CONTROL TREATMENT (B) INCENTIVE TREATMENT

Average reported confidence in the accuracy of response with 95% confidence intervals. Scale "1" Not at all confident to "7" Completely confident.

3 Main Results

3.1 Dishonesty in Response to Financial Incentives

We first ask whether the presence of financial incentives causes the participants to lie in their responses to the insurance claim. To answer this question, we pool together the responses to each of the 11 questions in the survey. We calculate the *change in the bonus* earned in the Base Incentive Treatment based on what the respondent added to or subtracted from their final payment with their responses. We calculate the bonus earned in the Control Treatment based on what the respondent *would have* added or subtracted from their final bonus, if there had been identical response-based incentives, with their responses. For example, if a participant's responses resulted in an added $10 on one response, an added $30 on another response, subtracted $10 on 3 responses, subtracted $20 on 2 other responses and had no additional changes on their bonus

for the remaining 4 questions, the net effect on their bonus would be subtracting $30 from their initial starting bonus of $350. Since the questions varied by how much could be added to the bonus with the responses, we also consider the percentage of the maximum possible payoff rendered by the subject's response. For example, if on one question the participant added $10 and the most they could have added was $40, then they added 25% of the maximum possible on this response, whereas if they added $10 and the maximum they could have added was $50 then they added only 20%.

In Table 3, we report OLS estimates of a model that regresses *Total $ of Lying* in column (1) and *Percent of Maximal Bonus* in column (2) on a dummy for the Base Incentive Treatment, using the Control Treatment as the omitted group. We find that, on average, subjects in the Base Incentive Treatment distort their answers to increase their bonus payment by $29.81 above the amount they would have received in the Control Treatment (i.e., without any financial incentives) and increase their percentage of maximal bonus by .43 percentage points (6% increase).

Table 3: Total Dishonesty in Response to Financial Incentives

	Total Lie in $	% of Maximal Bonus
Base Incentive Treatment	29.81***	0.43***
	(3.63)	(0.08)
Constant	-30.07***	7.40***
	(3.05)	(0.06)
Observations	459	459
R^2	0.14	0.07

OLS regression estimates. Robust standard errors in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.

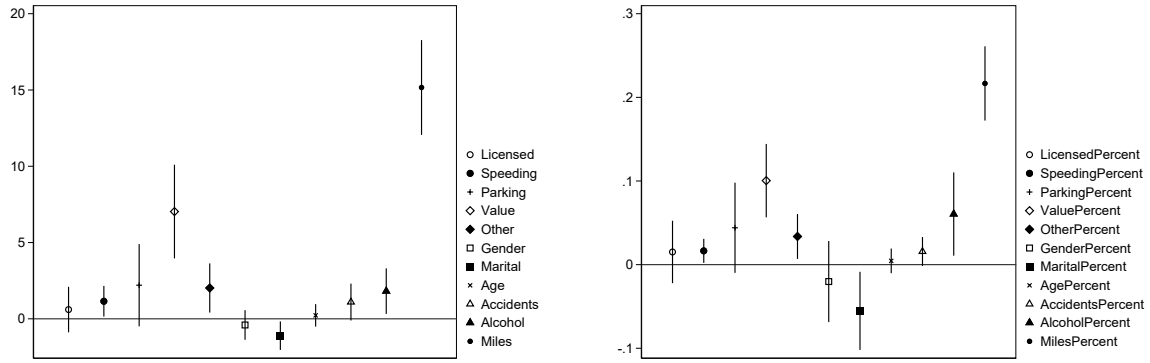
Result 1. *Participants' reports are affected by the presence of financial incentives to be dishonest. Participants who have an incentive to lie report answers to the 11-item survey such that they increase their bonus payment, on average, by $29.81 and increase their maximal bonus by 6% relative to those participants who have no incentive to lie.*

3.2 Self-Serving Truthfulness

Table 3 showed that subjects were dishonest in the presence of financial incentives. In this section, we provide evidence that this dishonesty is driven by self-serving truthfulness; that is, subjects distort their responses in a financially self-serving way, but only when they are uncertain of the objective truth. Recall, Figure 1 shows that subjects exhibit a high level of confidence in the correctness of their responses to each of the 11 survey questions, except for two questions, "Miles" and "Value". In fact, these are the only two questions that subjects give "below" average confidence in the correctness of their answers.

In Figure 2a, we plot the coefficients from 11 OLS regressions (1 regression for each question) that compare the average responses to each of the 11 survey questions in the Base Incentive Treatment relative to the Control Treatment. We find that 76% of the increase in the payoff found in Table 3 is driven by the responses given in the "Miles" and "Value" questions. Specifically, by under-reporting "Miles" and "Value", subjects in the Base Incentive Treatment, relative to subjects in the Control Treatment, earn on average a higher bonus of 15USD and 7USD, respectively.

Figure 2: Effect of Incentives on Dishonesty



(a) Dollar Effect of Dishonesty (b) Percentage of Dishonesty

OLS regression coefficients for each of the 11 questions in the survey. Figure 2a shows the dollar amount added to bonus in the Base Incentive Treatment relative to the non-incentivized Control Treatment. Figure 2b shows the percentage change in the dollar amount added to the bonus in the Base Incentive Treatment relative to the non-Incentivised Control Treatment. The bars represent 95% confidence intervals.

Table 4: Self-Serving Truthfulness

	Total Lie in $	% of Maximal Bonus	Total Lie in $	% of Maximal Bonus	Total Lie in $	% of Maximal Bonus
Incentive	11.10***	0.16***	0.85*	0.01	0.85*	0.01
	(1.19)	(0.02)	(0.45)	(0.01)	(0.47)	(0.01)
Incentive × Below Average Conf	.	.	.	.	10.25***	0.15***
					(1.11)	(0.02)
Below Average Conf	.	.	.	.	-33.93***	-0.13***
					(0.91)	(0.02)
Constant	-30.50***	0.56***	3.44***	0.7***	3.44***	0.7***
	(0.97)	(0.01)	(0.37)	(0.009)	(0.39)	(0.008)
Observations	918	918	4131	4131	5049	5049
	Below Average Confidence Questions		Above Average Confidence Questions		All Questions	

Random effects model using observations from the Base Incentive Treatment and Control Treatment. Robust standard errors clustered at the subject-level in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.

In Table 4 we estimate a random effects (mixed effects) model in which we pool together the responses for all 11 questions from subjects in the Control Treatment and the Base Incentive

Treatment, resulting in 11 observations per subject for 459 independent observations. We use the same two outcome variables as in Table 3. In columns (1) and (2), we compare the responses in the Base Incentive Treatment to the Control treatment only for those questions for which, on average, subjects report below average levels of confidence and we find that subjects distort their answers to increase their bonus by $11.10 and move 16 percentage points closer to the maximal possible bonus.

In columns (3) and (4), we estimate the same model as in columns (1) and (2) but only for those questions which subjects reported above average levels of confidence. Now we find minimal and insignificant effects of the Base Incentive Treatment.

In columns (5) (6), we pool together all 11 questions and interact the dummy for the incentive treatment with the dummy for the questions that have below average levels of confidence. Consistent with columns (1)-(4), subjects significantly distort their responses in the presence of financial incentives when they have below average confidence in the objectively correct response compared to when they have above average confidence.

Result 2. *Dishonesty in this experiment is driven by self-serving truthfulness; that is, when subjects are confident in the objective truth, there is very little evidence of dishonesty: in only two of the nine survey questions where subjects are confident in the objective truth do we detect lying, and this only accounts for 24% of the total gains from dishonesty. On the other hand, when subjects were less confident in the objective truth, we detect significantly more lying than in any of the questions where they knew the objective truth, and these two questions alone (just 18% of the questions) account for over three-fourths of all the financial gains received from dishonesty. Thus, when subjects do not know the objective truth, they are more likely to significantly and substantially distort their responses in a financially self-serving way than when they are certain of the objective truth.*

3.3 Self-serving truthfulness is non-strategic

Table 5: Testing Confidence Levels Between Control & Base Incentive Treatments

Licensed	Speeding	Parking	Value	Other	Gender	Marital	Age	Accidents	Alcohol	Miles
-0.62	2.63***	0.60	1.19	0.47	-1.06	-0.26	-1.06	-0.25	-0.84	-0.71

Z score test statistics from a Mann-Whitney test of differences where *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. We compare the distribution of responses about confidence in each of the 11 responses for the Control Treatment and the Base Incentive Treatment.

Last, we briefly examine whether financial incentives motivated subjects to self-deceive (Gneezy

et al., 2020); that is, we analyze whether subjects report lower confidence on their responses to questions in the presence of financial incentives in order to ex-post justify their lying. If this is the case, then we would expect to see less confidence in the Base Incentive Treatment than the Control Treatment. To test this, we compare the levels of confidence for each question between the Control and the Base Incentive conditions.

Table 5 reports the test statistic from a Mann-Whitney two-sample test for each of the 11 questions. We find that there are no significant differences in the confidence responses for 10 of the 11 questions, including critically both of the two questions with below average confidence (Miles and Value), between the Control Treatment (Figure 1a) and the Base Incentive Treatment (Figure 1b). Thus, we find no evidence that subjects were distorting their beliefs about the accuracy of their responses in a self-serving manner to justify their lying.

Result 3. *We find no evidence that self-serving truthfulness is strategic. Subjects do not report higher levels of uncertainty about the truthfulness of their responses when there is a financial incentive to lie.*

In sum, Result 1 and 2 provide evidence that when it is pay-off favorable, subjects distort their answers to the survey, but only for those questions in which they are significantly *less* certain of the accuracy of their response; that is, subjects display self-serving truthfulness. However, we find no evidence consistent with *strategic* self-serving truthfulness—subjects' confidence is unaffected by the presence of incentives to lie. This suggests that subjects respond to their exogenous level of uncertainty by "filling-in" their uncertainty in a self-serving way.[7]

3.4 Self-serving truthfulness is distinct from moral wiggle room

Moral wiggle room, like self-serving truthfulness, also stems from uncertainy. Self-serving truthfulness suggests that subjects distort their answers in a self-serving way when they are uncertain about the truth, but it is also possible for dishonesty to arise when individuals are uncertain about what is the morally appropriate behaviour. For example, individuals may argue that it is not clear what behavior is expected from them in a particular context. To mitigate dishonesty stemming from moral wiggle room with respect to rule clarity, we implement a set of Treatments including an honor statement that explicitly informs subjects that dishonesty is inappropriate and asks them to agree to behave honestly. We focus on the "Miles" and "Value" questions for which subjects are significantly less certain of their answers, since the vast majority of dishonesty stems from these two questions.[8]

[7]In Section 2, we also described a Treatment called the High Incentive Treatment in which we doubled the incentives offered relative to the Incentive Treatment. We find similar patterns of behaviour. First, subjects display behaviour consistent with dishonesty, but this is driven by the "Value" and "Miles" question (see Figure A1a) over which subjects are significantly less confident in the accuracies of their answers (see Figure A1b).

[8]Table A1 replicates Table 6 for the other 9 questions in the survey.

Result 4. *The interventions targeting moral wiggle room strategies reduce total dishonesty in the Miles question by $3.68 (24%) and are ineffective at mitigating dishonesty in the "Values" question. Of the moral wiggle room interventions, the Honor Code with the Check Box and the Honor Code with the Signature on the same screen were effective at decreasing dishonesty in the "Miles" question.*

Table 6: Self-Serving Truthfulness & Moral Wiggle Room

	Value	Value	Miles	Miles
Moral Wiggle Room Interventions	-0.13	.	-3.68***	.
	(1.08)		(1.14)	
Honor Code with Check Box	.	0.28	.	-3.98**
		(1.49)		(1.61)
Honor Code with Signature on Prev Screen	.	-0.22	.	-2.35
		(1.35)		(1.48)
Honor Code with Signature on Same Screen	.	-0.45	.	-4.71***
		(1.48)		(1.64)
Control Treatment	-7.03***	-7.03***	-15.16***	-15.16***
	(1.56)	(1.56)	(1.58)	(1.58)
Constant	-13.90***	-13.90***	-24.90***	-24.90***
	(0.84)	(0.84)	(0.85)	(0.85)
Observations	918	918	918	918
R^2	0.03	0.03	0.09	0.09
Omitted Group	Base Incentive Treatment			

OLS regression estimates. Columns (1) and (3) pool the moral wiggle room interventions, while columns (2) and (4) consider each of the three moral wiggle room interventions separately. Robust standard errors in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.

In Table 6 we report coefficients from OLS regressions in which we regressed responses to the "Value" and "Miles" question on a dummy for assignment to the Moral Wiggle Room Interventions and the Control Treatment using assignment to the Base Incentive Treatment as the omitted category. The coefficients for each variable represent the additional money added to their bonus (deducted from their overall premium) by their reports *relative to* the Base Incentive Treatment. In column (1) we find that the Moral Wiggle Room interventions, on average, do not significantly change the responses to the Value question relative to the Base Incentive Treatment. Similarly, in column (2) we disaggregate each of the Moral Wiggle Room interventions and find that each of the individual interventions added approximately $6.80-$7.30 to their bonus relative to the Control Treatment, but also that none of the individual interventions mitigated the dishonesty motivated by financial incentives. However, columns (3) and (4) show that the Moral Wiggle Room interventions had a significant impact on mitigating dishonesty motivated by financial incentives in the "Miles" question. Column (3) reports that dishonesty due to financial incentives is mitigated by $3.68 or 24% ($3.68/$15.16 of the total change in the bonuses due to the financial incentives) and column (4) shows that this was driven by the Honor Code with the Check Box Treatment and the Honor Code with a Signature Treatment.

3.5 Rational Lying & Consequence Attenuation

We also implemented interventions aimed at two additional strategies for justifying dishonesty: (1) Rational Lying and (2) Consequence Attenuation. We briefly discuss the results from these interventions, but we relegate the results to Tables B2 and B3 in Appendix B for two reasons. First, the interventions we used are not as developed in the literature as those from the moral wiggle room intervention and thus there may be specific interventions aimed at rational lying and consequence attenuation that we do not include here and that might yield different results. Second, our interventions were hypothetical rather than involving actual financial penalties, which is core to the rational lying hypothesis, and thus we interpret the results with caution.

Individuals who engage in consequence attenuation convince themselves that their behaviour is victimless and thus unharmful, or that the victim simply deserves to be harmed. We design two interventions aimed at these strategies: (1) Image Treatment and (2) Victim Treatment. The Image Treatment is designed to remind subjects that the insurance companies are not faceless profit-maximizing corporations but are organizations that also engage in socially beneficial endeavors. The implication is that acting dishonestly towards the insurance company could have negative repercussions on their ability to help others. The Victim Treatment reminds subjects that statements that falsely increase their claims will have negative externalities for other policy holders. Pooled together, we find that the Consequence Attenuation Intervention reduces dishonesty in the "Miles" question by \$2.72 (11%) but has no effect on responses to the "Value" question.

A final motivation for lying that we explore is rational dishonesty: if the cost of being punished for dishonesty is low, then individuals will engage in more dishonesty. To examine this motivation, we design the Rational Lying intervention, which reminds subjects that when claims are made the insurer may verify the information provided and if false information is detected then the individual may forfeit their right to any claim. We find that the Rational Lying Intervention has no effect on reducing dishonesty in the "Miles" or "Value" question.

4 Conclusion

We provide experimental evidence that questionnaire items asking for information that respondents may not be sure about are prone to trigger responses that are biased in respondents' monetary favour. In particular, we ask a sample of U.S. car owners to respond to a questionnaire that resembles an auto insurance underwriting questionnaire. When facing monetary incentives to indicate particular responses (reflecting the calculation of insurance premiums based on the information provided), we observe dishonesty, but almost entirely arising for the questions where participants are not sure about the correctness of the response they indicated. Participants are thus behaving

consistently with "self-serving truthfulness", meaning they respond to their uncertainty about the objectively correct response in a self-serving manner and choose responses that are more financially beneficial for them. Interventions that target commonly used strategies to justify dishonest behaviour like exploiting moral wiggle room with respect to rule clarity, attenuating the consequences of dishonesty for others, and attenuating the reprehensibility of accepting negative externalities, are only minimally effective in reducing dishonesty, indicating that self-serving truthfulness is a unique form of lying that has not been previously recognized.

References

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for truth-telling." *Econometrica* 87 (4):1115–1153.

Ayal, Shahar, Francesca Gino, Rachel Barkan, and Dan Ariely. 2015. "Three principles to REVISE people's unethical behavior." *Perspectives on Psychological Science* 10 (6):738–741.

Bandura, Albert. 1999. "Moral disengagement in the perpetration of inhumanities." *Personality and social psychology review* 3 (3):193–209.

Becker, Gary S. 1968. "Crime and punishment: An economic approach." In *The economic dimensions of crime*. Springer, 13–68.

Bellé, Nicola and Paola Cantarelli. 2017. "What causes unethical behavior? A meta-analysis to set an agenda for public administration research." *Public Administration Review* 77 (3):327–339.

Chou, Eileen Y. 2015a. "Paperless and soulless: E-signatures diminish the signer's presence and decrease acceptance." *Social Psychological and Personality Science* 6 (3):343–351.

———. 2015b. "What's in a name? The toll e-signatures take on individual honesty." *Journal of Experimental Social Psychology* 61:84–95.

Cressey, Donald R. 1986. "Why managers commit fraud." *Australian & New Zealand Journal of Criminology* 19 (4):195–209.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory* 33 (1):67–80.

Exley, Christine L. 2016. "Excusing selfishness in charitable giving: The role of risk." *The Review of Economic Studies* 83 (2):587–628.

Fukukawa, Kyoko. 2002. "Developing a framework for ethicallyquestionable behavior in consumption." *Journal of Business Ethics* 41:99–119.

Gneezy, Uri, Silvia Saccardo, Marta Serra-Garcia, and Roel van Veldhuizen. 2020. "Bribing the self." *Games and Economic Behavior* 120:311–324.

Grossman, Zachary and Joel J Van Der Weele. 2017. "Self-image and willful ignorance in social decisions." *Journal of the European Economic Association* 15 (1):173–217.

Haisley, Emily C and Roberto A Weber. 2010. "Self-serving interpretations of ambiguity in other-regarding behavior." *Games and economic behavior* 68 (2):614–625.

John, Leslie K, George Loewenstein, and Drazen Prelec. 2012. "Measuring the prevalence of questionable research practices with incentives for truth telling." *Psychological science* 23 (5):524–532.

Köneke, Vanessa, Horst Müller-Peters, Detlef Fetchenhauer et al. 2015. *Versicherungsbetrug verstehen und verhindern*. Springer.

Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in allocation decisions." *American economic review* 90 (4):1072–1091.

Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological bulletin* 108 (3):480.

List, John A, Sally Sadoff, and Mathis Wagner. 2011. "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design." *Experimental Economics* 14 (4):439.

Maggian, Valeria and Marie Claire Villeval. 2016. "Social preferences and lying aversion in children." *Experimental Economics* 19 (3):663–685.

Mazar, Nina, On Amir, and Dan Ariely. 2008. "The dishonesty of honest people: A theory of self-concept maintenance." *Journal of marketing research* 45 (6):633–644.

Rabin, Matthew. 1995. "Moral preferences, moral constraints, and self-serving biases." .

Schweitzer, Maurice E and Christopher K Hsee. 2002. "Stretching the truth: Elastic justification and motivated communication of uncertain information." *Journal of Risk and Uncertainty* 25 (2):185–201.

Shalvi, Shaul, Francesca Gino, Rachel Barkan, and Shahar Ayal. 2015. "Self-serving justifications: Doing wrong and feeling moral." *Current Directions in Psychological Science* 24 (2):125–130.

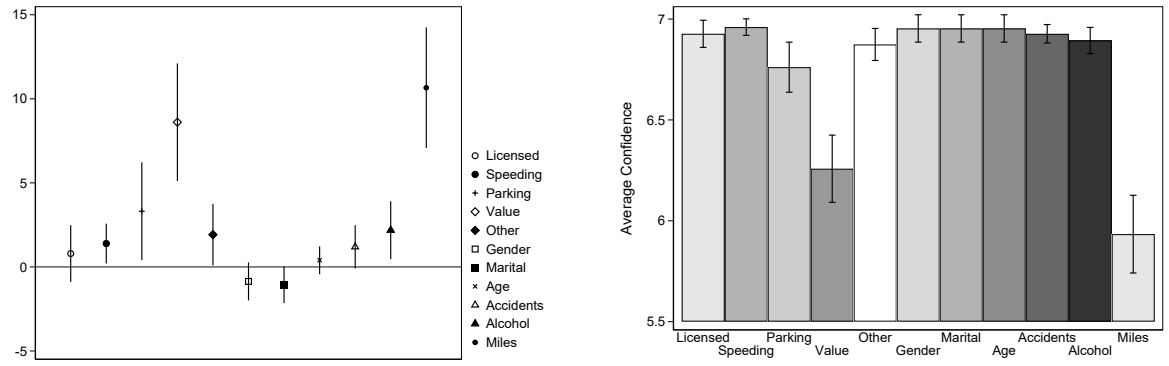
Shu, Lisa L, Francesca Gino, and Max H Bazerman. 2011. "Dishonest deed, clear conscience: When cheating leads to moral disengagement and motivated forgetting." *Personality and social psychology bulletin* 37 (3):330–349.

Shu, Lisa L, Nina Mazar, Francesca Gino, Dan Ariely, and Max H Bazerman. 2012. "Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end." *Proceedings of the National Academy of Sciences* 109 (38):15197–15200.

Sykes, Gresham M and David Matza. 1957. "Techniques of neutralization: A theory of delinquency." *American sociological review* 22 (6):664–670.

Appendix A Appendix

Figure A1: Data from the High Incentive Treatment



(a) Dishonesty (b) Confidence

Figure A1a plots coefficients from an OLS regression in which we compare reports from the High Incentive Treatment to the No Incentive Control. Figure A1b plots the reported confidence levels in the accuracies of the responses to each question in the 11-item survey. In both Figures, the bars represent 95% confidence intervals.

Table A1: Self-Serving Truthfulness & Moral Wiggle Room: All Questions

	Licensed	Speeding	Parking	Other	Gender	Marital	Age	Accidents	Alcohol
Moral Wiggle Room Interventions	1.87***	1.27**	3.48***	2.12***	-0.27	-0.61	0.26	1.22**	1.87***
	(0.69)	(0.5)	(1.26)	(0.8)	(0.47)	(0.44)	(0.35)	(0.6)	(0.72)
Incentive Treatment	0.61	1.15**	2.20	2.02**	-0.4	-1.11**	0.23	1.10*	1.81**
	(0.76)	(0.51)	(1.37)	(0.82)	(0.49)	(0.48)	(0.37)	(0.61)	(0.76)
Constant	-4.44***	-2.45***	31.85***	-3.77***	-4.24***	6.75***	-1.59***	-3.05***	11.85***
	(0.63)	(0.45)	(1.15)	(0.75)	(0.4)	(0.38)	(0.31)	(0.55)	(0.63)
Observations	918	918	918	918	918	918	918	918	918
R^2	0.01	0.009	0.01	0.01	0.0007	0.006	0.0007	0.007	0.008
Omitted Group	Control Treatment								
F tests									
Moral Wiggle Room = Incentive	6.27**	0.14	1.94	0.06	0.13	1.85	0.02	0.10	0.01

OLS regression estimates. Robust standard errors in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.

Appendix B Other Interventions

Table B2 shows the average effects of the Consequence Attenuation and Rational Lying Interventions at mitigating dishonesty. Table B3 we also examine the differences in responses that arise from each individual intervention Treatment relative to the Incentive using a Mann-Whitney test. In this table, we report the z-score associated with each of the tests and report the level of significance. In sum, none of the interventions are effective at mitigating self-serving truthfulness in the Value question, however a fraction of the dishonesty in the Miles question is mitigated by interventions targeting moral wiggle room and attenuation strategies. In particular, each of the honor codes implemented under the moral wiggle room interventions contributed to reducing dishonesty. On the other hand, of the attenuation interventions implemented, only the Victim Treatment, and not the Image Treatment, effectively mitigated dishonesty.

Table B2: Self-Serving Truthfulness, Attenuation & Rational Lying

	Licensed	Speeding	Parking	Value	Other	Gender	Marital	Age	Accidents	Alcohol	Miles
Attenuation Intervention	1.19**	0.26	1.64*	-0.37	0.52	-0.11	0.8**	0.09	-0.3	1.25**	-2.72**
	(0.56)	(0.32)	(0.98)	(1.20)	(0.4)	(0.4)	(0.39)	(0.28)	(0.41)	(0.59)	(1.26)
Rational Lying Intervention	0.61	0.71**	0.55	1.73	0.44	0.89*	0.6	0.25	0.17	0.21	-2.40
	(0.7)	(0.31)	(1.28)	(1.40)	(0.55)	(0.49)	(0.48)	(0.34)	(0.45)	(0.76)	(1.56)
Control Treatment	-0.61	-1.15**	-2.20	-7.03***	-2.02**	0.4	1.11**	-0.23	-1.10*	-1.81**	-15.16***
	(0.76)	(0.51)	(1.37)	(1.56)	(0.82)	(0.49)	(0.48)	(0.37)	(0.61)	(0.76)	(1.58)
Constant	-3.83***	-1.30***	34.06***	-13.90***	-1.75***	-4.64***	5.65***	-1.36***	-1.95***	13.67***	-24.90***
	(0.42)	(0.25)	(0.75)	(0.84)	(0.32)	(0.28)	(0.28)	(0.21)	(0.28)	(0.43)	(0.85)
Observations	918	918	918	918	918	918	918	918	918	918	918
R^2	0.009	0.02	0.01	0.03	0.02	0.005	0.007	0.002	0.006	0.02	0.1
Omitted Group	Base Incentive Treatment										

OLS regression estimates comparing the Attenuation Intervention & Rational Lying Intervention to the Base Incentive Treatment and Control Treatment. Robust standard errors in parentheses and *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively.

Table B3: Non-Parametric Test of the Interventions on Mitigating Dishonesty in the Value and Miles Question

	Moral Wiggle Room			Attenuation		Rational Lying
	Check Box	Signature	Signature PS	Image	Victim	Detection
Miles	-2.41**	2.57**	1.83*	1.08	2.16**	1.42
Value	0.42	0.76	0.31	0.56	0.03	-1.08

Z score test statistics from a Mann-Whitney test of differences where *, ** and *** indicate statistical significance at the 10%, 5% and 1% levels, respectively. We compare the distribution of responses from the Miles and Value question for each intervention Treatment to the Incentive Treatment.