

MOTIVATED SELF-DECEPTION, IDENTITY, AND UNETHICAL BEHAVIOR*

Uri Gneezy
Silvia Saccardo
Marta Serra-Garcia
Roel van Veldhuizen

We examine the role of self-deception in distorting judgment. We experimentally vary when evaluators are informed about incentives to recommend one of two options: before or after their initial private judgment. When the information regarding the incentives is provided before the judgment, we find a significant bias in the direction of the incentive. However, when the information is provided after they privately evaluate the options, but before they made their choice, participants' bias in judgment is significantly reduced. We term this behavior "motivated self-deception," arguing that in the before treatment judgment is biased such that evaluators can earn more money without compromising their self-image. Importantly, this bias in judgment occurs only when judgment is subjective and individuals are able to convince themselves that their behavior is ethical.

Keywords: Unethical Behavior, Motivated Self-Deception, Laboratory Experiment, Self-Image

JEL Classification: D03, D83, C91

* Gneezy: University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093 and University of Amsterdam (email: ugneezy@ucsd.edu). Saccardo: Carnegie Mellon University, Social and Decision Sciences, 5000 Forbes Avenue, BP 208, Pittsburgh, PA 15213 (email: ssaccardo@andrew.cmu.edu). Serra-Garcia: University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093 (email: mserragarcia@ucsd.edu). Van Veldhuizen: WZB Berlin Social Science Center, Reichpietschufer 50, 10785 Berlin, Germany (email: roel.vanveldhuizen@wzb.eu). This research was conducted under IRB#110022 and benefited from financial support of internal funds at Rady School of Management, UCSD. We thank participants in various seminars and conferences for their comments and suggestions.

I. Introduction

Unethical behavior, such as corruption or dishonesty, is widespread and comes with efficiency and fairness costs (Banerjee, 1997, and Svensson, 2005). For some people, distorting ethical judgment comes with a cost to self-image (Bem, 1972; Akerlof and Kranton, 2000; Bénabou and Tirole, 2006; Mazar, Amir and Ariely, 2008; Ariely, Bracha and Meier, 2009; Gneezy et al., 2012). All else equal, people who have such psychological costs prefer an outcome that is achieved without unethical choices to one that requires an action they consider unethical. To avoid this cost to self-image, people may choose actions that reduce their material payoffs. This conflict creates a tension between maintaining the self-image as a moral person and the desire to increase material goals. However, this tension may be attenuated if individuals can inadvertently convince themselves that their behavior is ethical.

Consider the healthcare sector, where overtreatment is estimated to cost \$210 billion in wasteful annual spending in the US (IOM, 2012), as well as obvious non-monetary costs to patients. One possible reason for overtreatment is that doctors frequently have financial incentives to recommend certain procedures for which they are directly compensated (Emanuel and Fuchs, 2008; Clemens and Gottlieb, 2014). Take for example the growing number of surgeries in response to back pain, many of which have been shown to be unnecessary and even harmful (Mafi et al., 2013). Another example is the large fraction of doctors who recommend unneeded C-sections for birth delivery when such procedures are financially compensated (see e.g., Gruber, Kim and Mayzlin, 1999; Johnson and Rehavi, 2015).

Some doctors may recommend unnecessary care knowingly in order to earn more money. Others, given that medical judgment is partially subjective, may convince themselves that the treatment they are prescribing is needed, thereby preserving their self-image as ethical professionals. In general, when judgment is subjective,

evaluators may unconsciously form their recommendation in a self-serving manner, while preserving their identity. We call this behavior “motivated self-deception,” where the decision maker can inadvertently convince herself that her behavior is ethical, preserving her identity as a moral person, while choosing the option that increases her personal gain. Motivated self-deception may also affect financial, legal, or policy advice, and, more generally, behavior in situations where individuals have room for maximizing their gains while preserving their identity as moral people.

The main question we ask in this paper is whether the evaluator knows that her evaluation is biased. That is, do evaluators distort their judgment knowingly, or do they engage in self-deception, convinced that their choice is ethical.

In the experiment we report in this paper, an advisor is asked to recommend one of two investment choices to a client. The two options differ in risk and expected return, there is no correct or incorrect choice, and the advisor receives a bonus if he recommends a specific investment option.

The key experimental manipulation contrasts two timelines of decision-making. In the first, the advisor is told about the incentives to choose one of the options *before* she is presented with the options she needs to consider. In the second, the advisor is told about the incentives only *after* seeing the two options and being asked to consider which one she would recommend. Importantly, in both cases, the advisor knows about the incentives before we observe her choice. That is, any judgment about the different options before providing the final recommendation only occurs in the advisors’ mind.

If the advisor is informed about the incentives before evaluating the options, she might be biased in her evaluation, without even realizing she is. If she first decided about her choice, and only then learns about the incentives, she might still recommend the option for which she is incentivized, but she would not be able to maintain the self-image of ethical choice.

Comparing these two basic manipulations allows us to identify the effect of a larger scope for self-deception on recommendations: Delaying the information regarding the incentives results in a significant reduction in the proportion of advisors who favor the incentivized option. Some people choose the incentivized option in any case, but a large portion does so only when able to convince themselves that they are not cheating. When the information about the incentive is delayed, individuals would still have room for rationalizing a recommendation of the incentivized option, as their initial evaluation only occurs in their mind, but they cannot fully believe that their advice is ethical. As we discuss below, previous literature has focused on ex-post rationalization of behavior and not on self-deception that occurs in the process of forming a judgment.

In an additional experiment, we further show that when the evaluation task is objective such that one investment option strictly dominates the other in every state of the world, and therefore advisors cannot convince themselves that the incentivized option is the ethical one, delaying the information about incentives has no effect on recommendations. This offers support to our argument that the difference in recommendations arising from a delay in the information regarding incentives in a subjective task is caused by self-deception.

Put together, our results support the hypothesis that incentives influence judgment to a much greater degree when evaluators are not aware that they are distorting their judgment for their own benefit, and hence reduce the cost to self-image. Our findings suggest a simple solution to some of the biased outcomes discussed above: separating the evaluation task from the information about incentives such that evaluations are formed before incentives can distort judgment.

Going back to the physicians' example, one solution to prevent overtreatment could be to inform physicians about incentives, e.g., details of their patients' insurance, only *after* they have a chance to evaluate which types of medical procedures are needed. Altering patient charts such that information about

insurance does not appear is easy to implement and could de-bias recommendations for at least the first visit.

Our results also suggest that reducing subjectivity and vagueness by asking physicians to delineate stricter criteria *ex ante* could be helpful in limiting the scope for self-deception. For those who consider themselves ethical but may fall prey to motivated self-deception, these interventions would prevent or lessen the extent of their unethical behavior.

In the next section we describe some of the related literature. Section III and IV describe the experimental procedure and results, and Section V concludes.

II. Related Literature

Our paper is motivated by the early work of Freud (1933) and Festinger (1957) on cognitive dissonance, and subsequent work on motivated reasoning (Lord, Ross and Lepper 1979; Kunda, 1990). This work suggests that individuals adjust their cognitions to reduce or eliminate conflicting desires in a variety of ways. Self-deception is defined as a situation in which an individual holds two contradictory beliefs without being aware of holding one of them, and such lack of awareness is motivated (Gur and Sackeim, 1979). It has been studied theoretically by, e.g., Benabou and Tirole (2002), Bodner and Prelec (2003), Brocas and Carrillo (2008), and Mijovich-Prelec and Prelec (2010). Existing empirical studies of self-deception, however, do not demonstrate lack of awareness.

Quattrone and Tversky (1984) required individuals to submerge their arm in a bucket of cold water and were told that high (low) tolerance was indicative of good health. In a second trial, individuals shifted their pain tolerance threshold according to what was indicative of good health. Such behavior is consistent with the idea that individuals denied to themselves that they changed their pain tolerance to form a

favorable diagnosis about their own state of health. However, there is no evidence of lack of awareness in this ex-post self-serving adjustment of behavior. Following Quattrone and Tversky (1984), studies have focused on biased *ex-post* rationalizations of own behavior or ability (see, e.g., Sloman, Fernbach, and Hagmayer, 2010; Chance et al., 2011; Fernbach, Hagmayer and Sloman, 2014). In this paper, we focus on self-deception that occurs *in the process* of judging a situation involving a conflict between material incentives and self-image about one's ethicality.

Our experimental procedure is related to the work by Babcock et al. (1995) on biased fairness judgments in negotiations (see also, Babcock and Loewenstein, 1997; and Konow, 2000 and Haisley and Weber, 2010 for such biases in allocation decisions). In Babcock et al. (1995), participants assigned to the role of plaintiff or defendant read a description of legal testimony. They were asked to indicate what they considered to be a fair outcome, and subsequently had to bargain over the actual settlement. Their design used a before-after manipulation like our experiments: In one treatment, participants were informed about their role as plaintiff or defendant before reading the legal testimony and making their fairness assessment, whereas in the other treatment they were informed only after. When participants learned their role before reading the case, fairness evaluations were more extreme, and settlements less likely.

Unlike our experiment, in their experiment participants in the before treatment had a strategic motive to find the best possible evidence supporting their case, and then use it in the negotiation. Given the strategic component, which is an important focus of their paper, it is not clear that individuals were unaware of forming self-serving assessments of the case. Our experiment focuses on how incentives and the desire to preserve the self-image as moral affect advice absent of any strategic considerations. Further, individuals in the after treatment in Babcock et al. (1995) were asked to provide their *written* assessment about fair outcomes before knowing

their roles in the negotiation, which makes this initial decision impossible to ignore. In our experiment, all initial assessments happen in people’s mind, thereby allowing us to attribute the change in behavior between the before and after treatment to self-deception.

Our paper relates to the literature on cheating (e.g., Mazar et al., 2008; Shalvi et al., 2011 and 2012). Differently from our experiment, in these studies individuals cheat about objective outcomes for which they have private information. Our results show that objective outcomes do not leave room for self-deception, which requires vagueness about the ethical choice. Further, we contribute to the literature on conflicts of interest in advice (e.g., Moore and Loewenstein, 2004; Cain, Loewenstein and Moore, 2005; Loewenstein, Cain and Sah, 2011; Sah, Loewenstein and Cain, 2013), focusing on how recommendations are affected by the tension between self-interest and identity as moral.

More broadly, our paper is related to work suggesting that individuals derive utility from forming positive beliefs about themselves (e.g. Köszegi, 2006), and to evidence suggesting that individuals actively avoid inconvenient information (e.g. Dana, Weber and Kuang, 2007; see Golman et al, 2015 for a review), and update beliefs asymmetrically in response to good and bad signals (Eil and Rao, 2012; Mobius et al., 2011). These phenomena could be considered tools individuals use to deceive themselves.

III. Distorted Advice Experiment

III.A. The Setting

In this experiment, we study a sender-receiver game in which the sender (“advisor”) is informed about the details of two investment opportunities, A and B, and is asked to send a recommendation to an uninformed receiver (the “client”)

regarding which of the two to choose. This game differs from standard sender-receiver games in that the sender is asked to make a judgment instead of reporting an objective piece of information, such as the state of nature (Crawford and Sobel, 1982).

The timeline of the experiment was as follows. First, the advisor was presented with information regarding the investment opportunities, A and B. Then she wrote a message recommending an option to a client. The client was a participant in a different experimental session and received no information about A and B. He only received the recommendation of the advisor and was asked to choose between A and B.

The information was presented to the advisors on four separate pages on their computer screen (all instructions are provided in the Appendix). On the first page, the advisor was informed about her role in the experiment and that she would be given a fixed payment of \$1 for participation. She was told that her role in the experiment would be to recommend one of two investment options (A and B) to another participant in a different session. She also learned that the other participant received no information about A or B except her recommendation.

On the second page of the instructions, advisors were presented with the details of A and B. The investment opportunities, labeled as product A and B, were described as having a 50% chance of being of high quality and a 50% chance of being of low quality. The payoff to the client for investment A was a 50-50 lottery between \$2 and \$4. Investment B was a 50-50 lottery between \$1 and \$7 dollars. The expected payoff of B (\$4) was higher than that of A (\$3). However, B had a higher variance. Thus, a tradeoff existed between risk and return across the two lotteries, such that the advisor could justify either choice by arguing (to herself) that risk or return was the more important factor for the recommendation.

In addition to receiving information about the lotteries, the advisor was asked (at the bottom of the screen) to think about her recommendation and continue to the

next screen once she was ready to provide it. Once the advisor moved to the third screen, the instructions asked her to raise her hand so that the research assistant could bring her the paper on which she would write her recommendation. Once she received the paper, she was asked to move onto the fourth and final screen, where she would provide her recommendation both on paper and on screen. This procedure allowed us to have the advisor send a message in her own handwriting, making the recommendation more tangible, as well as have a direct electronic record of recommendations.

The experiment had three treatments. In the Control treatment, advisors received no additional payment for recommending A or B. In the Before and After treatments, the advisor was told she would receive an additional commission of \$1 if she recommended A. The key difference between the Before and After treatments was *when* the advisor was first informed about the additional payment. In the Before treatment, advisors learned this information on the first screen, *before* learning the details of the investments. By contrast, in the After treatment, the advisor learned about the commission only on the fourth and final screen, *after* reading about the investments and having already thought about her recommendation, but before making the recommendation. To introduce only one change across treatments, the information on the commission was also presented on the fourth screen in the Before treatment.

If the only factor affecting which product the advisor recommends is the incentive, we should see no difference between the Before and After treatments. Assuming advisors are self-interested, and assuming they expect the client to follow their recommendation, they would recommend A in both treatments. If their self-image cost of distorting judgment is large enough, and the timeline of the experiment does not bias their evaluation of A and B, they would recommend A at the same rate in both treatments as in the Control treatment.

However, if the timing of the information about the incentive affects self-deception, whether the advisor knows about the commission of \$1 *before* or *after* reading about the investments may make a big difference. In the Before treatment, self-deception is easier, because the advisor learns about the incentives before seeing the products and may be able to convince herself that risk is undesirable, thereby giving her a reason to recommend A.

By contrast, in the After treatment, self-deception is harder. Here, the advisor has already made a decision about her evaluation of the tradeoff between risk and return before receiving information about incentives. Having initially decided to favor B, changing her recommendation to A may come at a cost to self-image, because she cannot deceive herself. Hence, motivated self-deception predicts that advisors will recommend A more often in the Before treatment than in the After and Control treatments.

III.B. Procedures

We conducted the experiment at the University of California, San Diego. Participants took part in an hour-long experimental session involving other studies. The experiment was run during two weeks and the order of studies in a session was the same within each week.¹ Randomization across the three treatments occurred at the participant level. As mentioned above, instructions were presented on computer screens and participants were asked to submit their recommendation for the client

¹ All other studies in a session were not incentivized and unrelated to our study. They were surveys in the fields of marketing and management, remained always the same and were presented in the same order within a week. We did not exclude subjects based on college major or past participation in other experiments. The only exclusion was that subjects must have not participated in this experiment before.

on a separate piece of paper, which only included the message “I recommend you to choose Product (A or B) _____.”²

We aimed at collecting 100 observations per treatment. Sessions were run for a whole day, and we stopped collecting data at the end of the day in which we achieved 300 observations. In total, 324 participants provided their recommendation as advisors (106 in Before, 110 in After, and 108 in Control). Forty-six percent of participants were female and the average age was 21.

One out of every ten recommendations was randomly selected and given to a client in a different session. Because the total number of recommendations was not a multiple of 10, we rounded it up and provided 33 recommendations to 33 clients.

A majority of the clients, 28 (84.8%) out of 33, followed the advisor’s recommendation. We found no difference in following depending on the recommendation, A or B (11 (91.7%) out of 12 and 17 (81%) out of 21, respectively; Fisher’s exact test, $p=.630$). Hence, the advisor’s recommendation had a high chance of directly affecting the client. In what follows, we focus on the behavior of advisors and examine the treatment effects on advisors’ recommendations.

We ran a second experiment in a different domain to examine the robustness of our results to a different setting in which there also is scope for self-deception. In this experiment (based on Gneezy, Saccardo and Van Veldhuizen, 2013), a referee is asked to award a prize to one of two workers according to their performance on a subjective real-effort task. Workers are given the opportunity to send money to the referee to influence her judgment. The same qualitative findings are obtained as in the main experiment and hence, for brevity, we report its results in Appendix A.

² Some participants (34 out of 324) did not follow the instructions as indicated. They did not raise their hand to request the paper for the message. We leave these participants in the sample and thereby report results conservatively. If we exclude these participants from the sample, results are strengthened in the direction of our prediction.

III.C. Results

Figure 1 displays the fraction of advisors recommending investment A in the three treatments. When information about the incentive tied to A is provided *before* reading about A, advisors are significantly more likely to recommend it. They recommend A in 43.4% of the cases in the Before treatment, compared to 27.7% of the cases in the After treatment. The difference is statistically significant (test of proportions, $Z\text{-stat}=2.481$, $p=.013$).

The percentage of advisors recommending A is 25.9% in the Control treatment and does not differ significantly from that in the After treatment ($Z=0.225$, $p=.822$). It does differ significantly from Before ($Z=2.687$, $p=.007$). Hence, we observe that the \$1 commission does not significantly distort judgment when announced after the information regarding the two lotteries; relative to control, the change is from 25.9% to 27.7%. However, it leads to a significant bias in recommendations when announced before the information on the lotteries, increasing A recommendations to 43.4%.

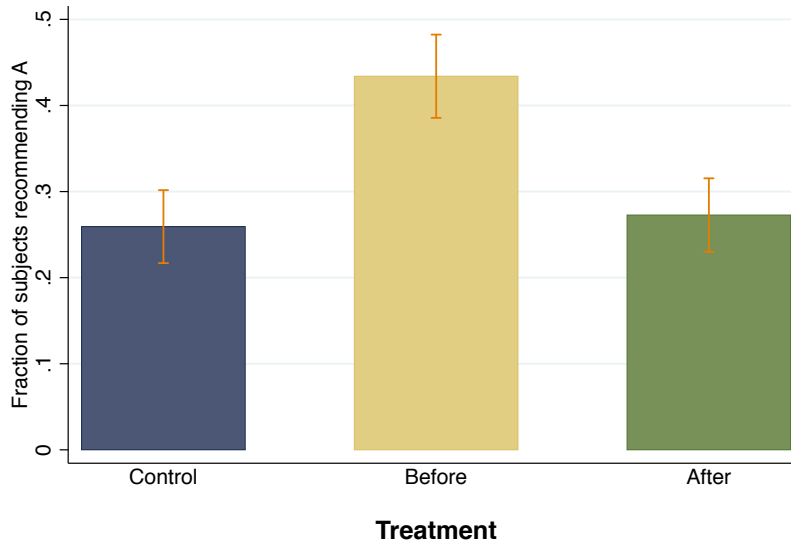


FIGURE 1: FRACTION OF ADVISORS RECOMMENDING A, BY TREATMENT

Notes: The figure presents the fraction of advisors who recommended option A in each of the three treatments respectively. The error bars represent ± 1 S.E.

Table 1 below confirms these results in a probit regression analysis. Column (1) confirms the average effect of Before relative to After: the likelihood of recommending A increases by 0.15 in the Before treatment, relative to the After treatment.³

We extend our analysis to examine the role of gender. Previous research has shown that females are more risk averse (e.g., Croson and Gneezy, 2009). In column (2), we introduce a control for gender, and in line with previous findings, we find that females are more likely to recommend A, which has a lower variance than B. In other words, female risk aversion appears to be reflected in female recommendations to others.

³ We conducted additional experiments, as will be described in what follows. In parallel, we conducted a Replication Experiment, in which we replicated this experiment to test its robustness to cohort effects in our subject pool. We recruited an additional 311 advisors following the same procedures (104 in the Control treatment, 103 in the Before treatment and 104 in the After treatment). We obtained even stronger treatment effects than in this experiment. There was a significant difference in A recommendations between the Before (60.2% of the cases) and the After (34.7% of the cases) treatments ($Z=3.826$, $p<.01$). In the Control treatment, A was recommended in 29.8% of the cases, which is not significantly different from the frequency of A recommendations in the After treatment ($Z=0.596$, $p=.551$).

Next, we examine whether the treatment effects vary by gender. Trivers (2011) suggests that men are more prone to self-deception than women. In the experiment, self-deception occurs only if an advisor would have recommended B in the absence of the incentive. Because a larger fraction of women recommend A in the Control treatment, the difference between Before and After may vary by gender. Columns (3) and (4) report the treatment effects splitting the sample by gender. We observe that whereas the effect of Before is strongly significant for men ($p < .001$), it is not significant for women ($p = .176$). However, the increase in A recommendations among men in the Before treatment, 16.5 percentage points, is not significantly different from the increase among women, 12.5 percentage points ($p = 0.319$). Thus, in the context of our experiment there is limited evidence of a gender difference in self-deception.

TABLE 1: TREATMENT EFFECTS ON THE LIKELIHOOD THAT A IS RECOMMENDED

	(1)	(2)	(3)	(4)
	All	P(A is recommended)		Female
		All	Male	
Before Treatment	.146*** (.051)	.159*** (.049)	.165*** (.044)	.125 (.093)
Control Treatment	-.014 (.061)	-.007 (.059)	-.009 (.071)	-.002 (.095)
Female		.178*** (.045)		
Share recommending A in After Treatment	.273	.273	.164	.382
Observations	324	324	174	150

Notes: Columns (1) to (4) report marginal effects from probit regressions on the likelihood that A is recommended. In columns (1) and (2), all advisors are included, whereas column (3) reports results only for male advisors and column (4) only for female advisors. The variables ‘Before Treatment’ and ‘Control Treatment’ are dummy variables taking value 1 if the treatment is Before or Control, respectively. The omitted category is the After treatment. Female is a dummy variable that takes value 1 if the participant is a female. Marginal effects are evaluated for a man (column 2) in the After treatment (columns 1 to 4). Standard errors are reported in parenthesis.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Our results support the prediction that providing incentives to recommend A leads to a stronger bias towards this option when the information about the incentive

is revealed *before* the advisor evaluates the two options (A and B) than when it is revealed *after* the options have been privately evaluated. This suggests that motivated self-deception may indeed have been harder in treatment After. When advisors were informed about the incentives before evaluation, motivated self-deception may have allowed them to color their judgment in the direction of the incentives. By contrast, when they were informed about the incentives after the initial evaluation, judgment was less biased.

Our results could also be consistent with two alternative explanations. One alternative explanation is that participants in treatment Before may have avoided evaluation altogether and simply recommended the incentivized option, either because of the incentives per se, or because they perceived the incentives as a signal that the incentivized option was in fact the better product. This would imply that participants in treatment Before require less time to finish the experiment. However, we do not find a significant difference in the time taken to complete the experiment between the Before and After treatments (Mann-Whitney test, $p=.170$), or relative to Control (Mann-Whitney test, Before vs. Control, $p=.215$; After vs. Control, $p=.829$). Second, the smaller bias could also result from preferences for consistency (see Cialdini, 1984) according to which advisors in the After treatment might have a preference to stick to the first judgment they formulated in their minds. We provide further evidence in support of self-deception, and against these alternative explanations, in an additional experiment in which we remove any scope for self-deception.

IV. Limiting the scope for motivated self-deception

According to our prediction, motivated self-deception occurs only when judgment is subjective. When evaluation occurs on multiple dimensions, such as

risk and return, and no option strictly dominates the other, individuals can choose the dimension they consider most relevant. Given that such choices are subjective, there is scope for participants to convince themselves that the dimension that is materially more advantageous to them is the most important. If instead an option *strictly* dominates the other in all dimensions we expect any scope for self-deception to be eliminated. This idea is in line with research in psychology on self-deception about personal abilities, which has shown that self-deception only occurs in situations characterized by ambiguity, vagueness or uncertainty, as they leave room for self-serving interpretations (see e.g., Sloman, Fernbach and Haggmayer, 2010; Fernbach, Haggmayer and Sloman, 2014).

In this section we present an experiment in which we introduce *strict* dominance in all dimensions, removing any scope for motivated self-deception. We predict that in such setting, the timing of the information regarding the incentives will not differentially affect choices.

IV.A. Strict Dominance Experiment

In an additional experiment, we introduced strict dominance between investments A and B. The only change relative to the previous experiment was the value of B: a 50-50 lottery between \$5 and \$7, instead of \$1 and \$7. Investment A remained unchanged—a 50-50 lottery between \$2 and \$4. Thus, in this experiment investment B yields a strictly better outcome than A in every state of the world.

As in the Distorted Advice Experiment, there were three treatments. In the Control treatment, there was no additional incentive for recommending A or B. In the Before and After treatments, advisors received a commission of \$1 for recommending A. Advisors were informed about the commission either *before* or *after* evaluating A and B.

Introducing strict dominance in the experiment removes the scope for motivated self-deception, since advisors cannot any longer convince themselves that A is the better option, as B strictly dominates A in every state of the world. Therefore, we predict no difference between Before and After in this experiment.

Importantly, the prediction of no difference between Before and After in this experiment also allows us to address the two alternative explanations discussed above. First, if the difference between Before and After is driven by participants avoiding evaluation in Before, we would still expect a higher rate of A recommendations in this treatment than in the After treatment. Second, if preferences for consistency would explain the lower rate of A recommendations in After because individuals stick with the judgment formed before learning about the incentive, we would also still expect a difference in recommendations between Before and After.

The procedures followed in this experiment were the same as in the Distorted Advice Experiment. We recruited 334 participants who provided their recommendation as advisors (113 in Control, 109 in Before, and 112 in After). Fifty-four percent of participants were female and the average age was 21.

A majority of the clients, 25 (73.5%) out of 34, followed the advisor's recommendation. We found no difference in following depending on the recommendation, A or B (8 (80%) out of 10 and 17 (70.8%) out of 24, respectively; Fisher's exact test, $p=.692$). Hence, as in the other experiment, the advisor's recommendation had a high chance of directly affecting the client.

IV.B. Results

Figure 2 displays the fraction of advisors recommending A in each treatment. In the Control treatment, where advisors do not receive any incentive for

recommending A or B, 15.9% recommend A. When an incentive to recommend A is introduced, the rate of A recommendations increases to 31.2% in the Before treatment. Importantly, the fraction A recommendations observed in the Before treatment is not significantly different from the fraction of (30.4%) observed in the After treatment ($Z=0.1345$, $p=.893$).

The presence of the incentive significantly increases A recommendations in both Before and After, relative to Control ($Z=2.685$, $p=.007$ comparing Before and Control; $Z=2.567$, $p=0.010$ comparing After and Control).

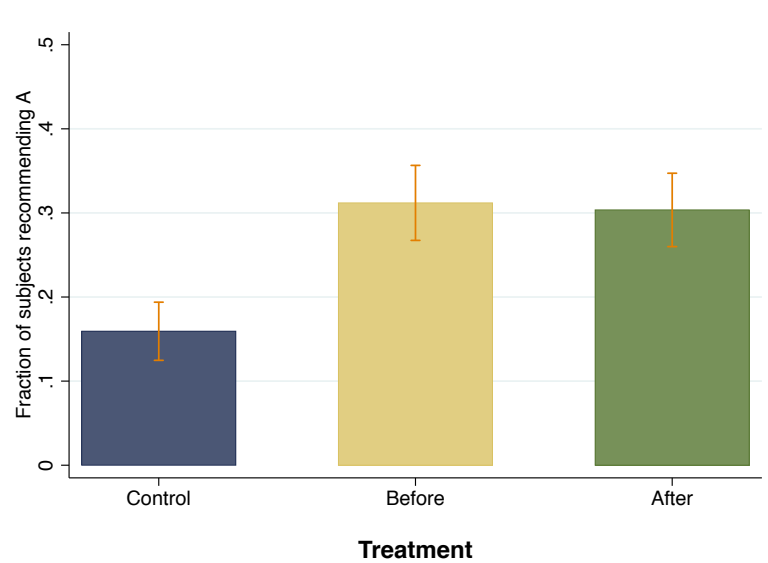


FIGURE 2: FRACTION OF ADVISORS RECOMMENDING A IN THE STRICT DOMINANCE EXPERIMENT, BY TREATMENT

Notes: The figure presents the fraction of advisors who recommended option A in each of the three treatments respectively. The error bars represent +/- 1 S.E.

Table 2 confirms the results using a probit regression analysis. Column (1) shows that there is no statistically significant difference between A recommendations in Before and After. Further, the magnitude of the marginal effect is very small, 0.008, in line with the difference in frequencies observed in Figure 2. The rate of A recommendations in the Control treatment is significantly higher than in the After

treatment ($p=.022$), as is the difference between the coefficients for the Before and Control treatments ($p<.01$).

Examining the role of gender, we find that there are no significant differences in A recommendations between female and male participants, as shown in column (2). This is in line with the strict dominance of B, which does not yield a risk-return tradeoff that could lead to different recommendations depending on the individual's degree of risk aversion, the explanation for the gender difference observed in the Distorted Advice Experiment. In columns (3) and (4) of Table 2 we examine the effects of the Before and Control treatment by gender. We do not find significant gender differences in the effect of the Before treatment ($p=.826$), or the Control treatment ($p=.897$).

TABLE 2: TREATMENT EFFECTS ON THE LIKELIHOOD THAT A IS RECOMMENDED IN THE STRICT DOMINANCE EXPERIMENT

	(1)	(2)	(3)	(4)
	All	All	Male	Female
Before Treatment	.008 (.061)	.009 (.061)	.024 (.089)	-.003 (.085)
Control Treatment	-.169** (.074)	-.168** (.074)	-.179* (.108)	-.159 (.100)
Female		-.016 (.053)		
Share recommending A in After Treatment	.304	.304	.309	.298
Observations	334	334	154	180

Notes: Columns (1) to (4) report marginal effects from probit regressions on the likelihood that A is recommended. In columns (1) and (2), all advisors are included, whereas column (3) reports results only for male advisors and column (4) only for female advisors. The variables 'Before Treatment' and 'Control Treatment' are dummy variables taking value 1 if the treatment is Before or Control, respectively. The omitted category is the After treatment. Female is a dummy variable that takes value 1 if the participant is a female. Marginal effects are evaluated for a man (column 2) in the After treatment (columns 1 to 4). Standard errors are reported in parentheses.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

The results of this experiment provide further evidence in support of the presence of motivated self-deception when judgment is subjective. Removing the scope for motivated self-deception by introducing a strict dominance relationship between the items to be judged removes any difference in recommendations when

information about incentives is delayed. This suggests that our original treatment effect was not due to the avoidance of evaluation or a preference for consistency. The latter result is in line with Falk and Zimmermann (2015), who show that individuals exhibit preferences for consistency only when they formulate their first judgment in writing, not when they do so in their mind, as in our experiments.

IV.C. The Persistence of Motivated Self-Deception: Weakening Dominance

The results thus far provide evidence for motivated self-deception when evaluation is performed on multiple dimensions and no option strictly dominates others in all dimensions. Yet, if *strict* dominance is introduced, no evidence of motivated self-deception is found. In this section we present an intermediate case: an experiment with *weak*, rather than strict dominance, providing a test of the persistence of motivated self-deception.

In the Weak Dominance experiment we again changed the payoffs associated with investment B. In this case, B was a 50-50 lottery between \$2 and \$6. A remained a 50-50 lottery between \$2 and \$4. There are two competing hypotheses. On the one hand, weak dominance could limit the scope for motivated self-deception in the same way as strict dominance does, since investment B weakly dominates investment A. On the other hand, previous findings suggest that even a minor reason to favor the incentivized option could be used by individuals to convince themselves that recommending that option is ethical (Kunda, 1990, see also, Konow, 2000). Hence, if the advisor focused on the “bad” state for both lotteries, or if she compared the increase in her payoff (\$1) to the decrease in the expected payoff of the recipient (also \$1), she could find reasons to recommend A.

We ran the Weak Dominance Experiment following the same procedures as in the experiments above. There were 300 advisors (100 in the Control treatment, 101 in the Before treatment and 99 in the After treatment).

The results of the Weak Dominance Experiment reveal that, when B only weakly dominates A, there is still a significant difference between the Before and After treatments. In the Before treatment, advisors recommended A in 53.5% of the cases, while in the After treatment, they recommended A in 25.3% of the cases ($Z=4.081$, $p<.01$). In the Control treatment, participants recommended A in 14% of the cases. This frequency was significantly lower than in the Before treatment ($Z=5.913$, $p<.01$) and than in the After treatment ($Z=1.999$, $p=0.046$).

The results suggest that motivated self-deception can be persistent. As long as there is weak dominance on some dimension upon which several items are evaluated, individuals seem able to focus on that dimension, ignoring other dimensions, and recommend the incentivized product without a cost to their self-image (in line with several studies cited in Kunda, 1990). However, when strict dominance is introduced, the bias introduced by incentives through motivated self-deception vanishes entirely.

V. Conclusion

Understanding why people behave unethically can help structuring policies to reduce such behavior. For example, many physicians believe incentives such as receiving fees for each procedure they perform or gifts from pharmaceutical companies do not influence their judgment. This belief allows them to receive the incentives while maintaining their self-image as unbiased physicians. The evidence suggests the physicians are wrong, and incentives do distort their judgment in many cases (Steinman et al., 2001; Cain, Loewenstein, and Moore, 2005 and 2011;

Malmendier and Schmidt, 2012). This biased judgment comes at a cost to the patients who may not receive the best available treatment and/or may pay more for it.

Examples in which ethical choices are biased by incentives are plentiful and have a huge impact on efficiency and fairness. How can policy makers change this practice? One clear way is to outlaw such incentives when possible, and enforce these laws. But in some cases, changing the law (e.g., due to lobbyists) or monitoring behavior (e.g., because judgment is subjective) can be hard. Even when this type of solution is feasible, enforcing it could be very costly.

In this paper, we propose an additional approach to reducing the effectiveness of incentives in distorting judgment. By having decision makers first evaluate the options and only then receive information about the incentives, we made them face their biased choices, changing the behavior of a significant fraction of our participants. We argue that this reduction in unethical behavior results from the psychological cost to the self-image: when faced with the bias, the decision maker cannot engage in motivated self-deception, convincing him/herself that the choices are ethical.

Our message is clear. Some people have psychological costs associated with distorting judgment. Creating procedures that reinforce the role of these psychological costs can reduce unethical behavior by ethical-but-biased individuals. In addition to healthcare and financial, legal and policy advice, our findings speak to the recent discussion in academia around the failure to replicate many published findings. Even though instances of data fabrication are part of the problem, another reason for this crisis is researchers who engage in questionable research practices that increase the chance of false positives (e.g. Simmons, Nelson, and Simonsohn, 2011; Gelman, 2013). Consider for example making “predictions” or choosing which analysis to perform only after looking at the data. Such degrees of freedom in the research practices may allow researchers to get the significant results needed

to publish their papers but at the same time feel as if they did not break any ethical rule, preserving their self-image.

One of the suggested solutions for this crisis was to create a clearinghouse to which researchers will have to submit the details of their design and the planned analysis before running the experiment. According to this suggestion, only experiments that were first registered will be considered for publication.

See for example the AEA registry of RCTs in economics and other social sciences (<https://www.socialscienceregistry.org/site/about>). This solution is hard to implement for laboratory experiments, and will surely come at a cost to piloting and being creative (see Gelman, 2013). An alternative, arguably simpler, approach is to encourage researchers to pre-specify to themselves the data-collection process (e.g., how many observations they plan to collect) and the analysis they plan to perform. Although reducing the degrees of freedom in research this way will not help reduce outright fraud, it might reduce the unethical behavior by people who consider themselves ethical.

REFERENCES

- Akerlof, George A., and Rachel E. Kranton. 2000. "Economics and identity." *Quarterly Journal of Economics* 115 (3): 715-753.
- Ariely, Dan, Anat Bracha, and Stephan Meier. 2009. "Doing good or doing well? Image motivation and monetary incentives in behaving prosocially." *American Economic Review* 99 (1): 544-55.
- Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased judgments of fairness in bargaining." *American Economic Review* 85 (5): 1337-1343.
- Babcock, Linda, and George Loewenstein. 1997. "Explaining bargaining impasse: The role of self-serving biases." *Journal of Economic Perspectives* 11 (1): 109-26.
- Banerjee, Abhijit V. 1997. "A theory of misgovernance." *Quarterly Journal of Economics* 112 (4): 1289-1332.
- Bem, Daryl J. 1972. "Self-perception theory." *Advances in Experimental Social Psychology* 6: 1-62.
- Bénabou, Roland, and Jean Tirole. 2002. "Self-confidence and personal motivation." *Quarterly Journal of Economics* 117 (3): 871-915.
- Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review* 96 (5): 1652-78.
- Bodner, Ronit, and Drazen Prelec. 2003. "Self-signaling and diagnostic utility in everyday decision making". In Isabelle Brocas and Juan D. Carrillo (Eds.), *The Psychology of Economic Decisions (1)* (pp. 105-26). Oxford University Press.
- Brocas, Isabelle, and Juan D. Carrillo. 2008. "The brain as a hierarchical organization." *The American Economic Review* 98 (4): 1312-1346.
- Cain, Daylian M., George Loewenstein, and Don A. Moore. 2005. "The dirt on coming clean: perverse effects of disclosing conflicts of interest." *Journal of Legal Studies* 34 (1): 1-25.
- Cain, Daylian M., George Loewenstein, and Don A. Moore. 2011. "When sunlight fails to disinfect: understanding the perverse effects of disclosing conflicts of interest." *Journal of Consumer Research* 37 (5): 836-57.
- Chance, Zoë, Michael I. Norton, Francesca Gino, and Dan Ariely. 2011. "Temporal view of the costs and benefits of self-deception." *Proceedings of the National Academy of Sciences* 108 (3): 15655-15659.
- Cialdini, Robert. 1984. *Influence, the Psychology of Persuasion*. New York: Harper Collins.
- Clemens, Jeffrey, and Joshua D. Gottlieb. 2014. "Do physicians' financial incentives affect medical treatment and patient health?" *American Economic Review* 104 (5): 1320-1349.

- Crawford, Vincent P., and Joel Sobel. 1982. "Strategic information transmission." *Econometrica* 50 (6): 1431-51.
- 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.
- Eil, David, and Justin M. Rao. 2011. "The good news-bad news effect: asymmetric processing of objective information about yourself." *American Economic Journal: Microeconomics* 3(2): 114-138.
- Emanuel, Ezekiel J., and Victor R. Fuchs. 2008. "The perfect storm of overutilization." *JAMA: The Journal of the American Medical Association* 299 (23): 2789-2791.
- Falk, Armin, and Florian Zimmermann. 2015. "Information Processing and Commitment." *Mimeo*.
- Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Evanston, IL: Row, Peterson.
- Fernbach, Philip M., York Hagmayer, and Steven A. Sloman. 2014. "Effort denial in self-deception." *Organizational Behavior and Human Decision Processes* 123(1): 1-8.
- Fischbacher, Urs, and Franziska Föllmi-Heusi. 2013. "Lies in Disguise-An Experimental Study on Cheating." *Journal of the European Economic Association* 11 (3): 525-47.
- Freud, Sigmund. 1933. *New Introductory Lectures on Psycho-Analysis*. W.W. Norton & Company. The Standard Edition edition (1990).
- Gelman, Andrew. 2013. "Preregistration of studies and mock reports." *Political Analysis* 21: 40-41.
- Gneezy, Ayelet, Uri Gneezy, Gerhard Riener, and Leif D. Nelson. 2012. "Pay-what-you-want, identity, and self-signaling in markets." *Proceedings of the National Academy of Sciences* 109 (19): 7236-40.
- Gneezy, Uri. 2005. "Deception: the role of consequences." *American Economic Review* 95 (1): 384-394.
- Gneezy, Uri, Silvia Saccardo, and Roel van Veldhuizen. 2013. "Bribery: greed versus reciprocity." *Working paper*.
- Golman, Russell, Hagmann, David and Loewenstein, George. 2016. "Information Avoidance." *Journal of Economic Literature, Forthcoming*. Available at SSRN 2633226.
- Gruber, Jonathan, John Kim, and Dina Mayzlin. 1999. "Physician fees and procedure intensity: the case of cesarean delivery." *Journal of Health Economics* 18 (4): 473-490.
- Gur, Ruben C., and Harold A. Sackeim. 1979. "Self-deception: A concept in search of a phenomenon." *Journal of Personality and Social Psychology* 37 (2): 147-169.

- Haisley, Emily C., and Roberto A. Weber. 2010. "Self-serving interpretations of ambiguity in other-regarding behavior." *Games and Economic Behavior* 68(20): 614-625.
- IOM (Institute of Medicine). 2012. *Best care at lower cost: The path to continuously learning health care in America*. Washington, DC: The National Academies Press.
- Johnson, Erin M., and M. Marit Rehani. 2015. "Physicians treating physicians: information and incentives in childbirth." *National Bureau of Economic Research Working Paper* No. 19242.
- Konow, James. 2000. "Fair shares: Accountability and cognitive dissonance in Allocation Decisions." *American Economic Review* 90 (4), 1072-1091.
- Kőszegi, Botond. 2006. "Ego utility, overconfidence, and task choice." *Journal of the European Economic Association* 4(4): 673-707.
- Kunda, Ziva. 1990. "The case for motivated reasoning." *Psychological Bulletin* 108 (3): 480.
- Loewenstein, George, Daylian M. Cain, and Sunita Sah. 2011. "The limits of transparency: Pitfalls and potential of disclosing conflicts of interest." *The American Economic Review* 101 (3): 423-428.
- Lord, Charles G., Lee Ross, and Mark R. Lepper. 1979. "Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence." *Journal of Personality and Social Psychology* 37 (11): 2098.
- Mafi, John N., Ellen P. McCarthy, Roger B. Davis, and Bruce E. Landon. 2013. "Worsening trends in the management and treatment of back pain." *JAMA internal medicine* 173 (17): 1573-1581.
- Malmendier, Ulrike, and Klaus Schmidt. 2012. "You owe me." *NBER Working Paper* No. 18543.
- 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.
- Mijović-Prelec, Danica, and Drazen Prelec. 2010. "Self-deception as self-signalling: a model and experimental evidence." *Philosophical Transactions of the Royal Society B: Biological Sciences* 365(1538): 227-240.
- Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. 2011. "Managing self-confidence: Theory and experimental evidence." *No. w17014. National Bureau of Economic Research*.
- Moore, Don A., and George Loewenstein. 2004. "Self-interest, automaticity, and the psychology of conflict of interest." *Social Justice Research* 17 (2): 189-202.
- Quattrone, George A., and Amos Tversky. 1984. "Causal versus diagnostic contingencies: On self-deception and on the voter's illusion." *Journal of personality and social psychology* 46 (2): 237.
- Sah, Sunita, George Loewenstein, and Daylian M. Cain. 2013. "The burden of

- disclosure: increased compliance with distrusted advice." *Journal of personality and social psychology* 104 (2): 289.
- Shalvi, Shaul, Jason Dana, Michel JJ Handgraaf, and Carsten KW De Dreu. 2011. "Justified ethicality: Observing desired counterfactuals modifies ethical perceptions and behavior." *Organizational Behavior and Human Decision Processes* 115 (2): 181-190.
- Shalvi, Shaul, Ori Eldar, and Yoella Bereby-Meyer. 2012. "Honesty requires time (and lack of justifications)." *Psychological science* 23 (10): 1264-1270.
- Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological Science* 22 (11): 1359-66.
- Sloman, Steven A., Philip M. Fernbach, and York Hagmayer. 2010. "Self-deception requires vagueness." *Cognition* 115(2): 268-281.
- Steinman, Michael A., Michael G. Shlipak, and Stephen J. McPhee. 2001. "Of principles and pens: attitudes and practices of medicine housestaff toward pharmaceutical industry promotions." *American Journal of Medicine* 110 (7): 551-7.
- Svensson, Jakob. 2005. "Eight questions about corruption". *Journal of Economic Perspectives* 19 (3): 19-42.
- Stroop, J. R. 1935. "Studies of interference in serial verbal reactions." *Journal of Experimental Psychology* 18 (6): 643-662.
- Trivers, Robert. 2011. *The folly of fools: the logic of deceit and self-deception in human Life*. Basic Books.

For Online Publication

Appendix A: Design and Results in the Distorted Choice Experiment

A.1. The Game

The distorted choice game (Gneezy, Saccardo and Van Veldhuizen, 2013) involves three players: two workers and a referee. The workers compete against each other in a real-effort task. The third player, the referee, is asked to judge the tasks and select the winner, who gets a prize of p . Each worker i is allowed to send an amount of money ($m_i \in [0, \frac{1}{2}p]$) to the referee, with only integer amounts allowed in the experiment. The referee can only keep the money of the worker who wins the prize.

The referees' payoff-maximizing strategy in this game is to choose as the winner the worker who sends the highest amount of money. Instead, if referees have moral costs associated with lying about who was the best performer (e.g., Gneezy, 2005), they will prefer to award the prize to the best performer of the real-effort task and will be willing to forgo some monetary benefit by doing so.⁴

However, even ethical referees may bias their judgment of the real-effort task to favor the worker who sent the highest amount of money. Such motivated self-deception could occur, for instance, if referees are able to convince themselves that the worker who sent the highest amount also performed better on the real-effort task, even if she in fact performed worse. If motivated self-deception is successful, referees can thus avoid the self-image cost associated with choosing the worst performance.

⁴ Gneezy, Saccardo, and Van Veldhuizen (2013) use this game to study the relative importance of greed and reciprocity in accepting bribes. Their key comparison is between a treatment in which referees can keep only the money sent by the winner and a treatment in which they keep the money from both workers. The main finding is that sending money is significantly more effective in the former than in the latter.

To investigate how motivated self-deception affects the referee's judgment, we ran three treatments that share the same payoff structure but differ in their scope for motivated self-deception. In the Before and After treatments, we asked referees to evaluate workers' performance on a subjective real-effort task that consisted of writing a joke about a pre-specified topic. Though some jokes are clearly better than others, humor is at least partially a matter of taste. As a result, we expected motivated self-deception to be relatively easy in this task.

As in the Distorted Advice Experiment, our main manipulation contrasts two timelines of decision-making. In the Before treatment, the referee received the jokes and the money sent by the workers simultaneously and was then asked to select the winner. Therefore, referees in this treatment had a chance to see the money sent *before* making their judgment about the quality of jokes. As a result, we expect referees to be able to engage in motivated self-deception, convincing themselves that the joke that corresponds to the highest amount sent is also the best joke. Thus, we predict that the choices of the referee in this treatment will favor the worker who sent the highest amount of money, regardless of the quality of the jokes.

In the After treatment, the referee received the money sent by the workers two minutes *after* receiving the jokes. For the first two minutes, the referee had a chance to evaluate the joke without being influenced by the incentives. Hence, the referee could form an unbiased judgment of the jokes before she received the incentives, and convincing herself that the worker who sent the higher amount of money was also the one with the best joke would have become more difficult. Choosing in favor of the worker who sent the highest amount of money is therefore likely to generate higher self-image costs than doing so in the Before treatment. Thus, we predict that incentives will play a smaller role and the quality of the jokes will play a larger role in this treatment.

We also ran a third treatment, "Objective," as an alternative test of our hypothesis. In this treatment, referees had to judge workers' performance on an objective real-

effort task. In particular, workers were asked to identify the colors of a sequence of words (Stroop, 1935). As in the Before treatment, referees in this treatment received the task output and the money sent by the workers simultaneously. However, because workers' performance was objective, engaging in motivated self-deception and appearing ethical to oneself is harder. Therefore, we predict that referees will select the worker with the best performance more often than in the Before treatment.

A.2. Procedures

We conducted the experiment at the University of California San Diego with 273 total participants, 6 in each session.^{5,6} Among the participants, 56% were female and the average age was approximately 21.

Upon arrival, we randomly assigned participants to computer terminals and provided them the instructions on computer screens. Participants were anonymously matched in groups of three and were assigned to the role of worker or referee. Each referee was then seated in a separate room and received a \$5 show-up fee. Each worker received a \$10 show-up fee in \$1 bills.

In the Before and After treatments, participants had 10 minutes to type a joke. The topic of the joke was "Economists," and it was communicated immediately before the beginning of the task. After they typed their jokes, workers were asked to report how confident they were that their joke was better than their competitor's. Each joke was then printed on a sheet of paper. Afterwards, workers were informed

⁵ The data of 123 participants (60 in Before, 63 in Objective) are also reported in Gneezy, Saccardo, and Van Veldhuizen (2013). Because we wanted to have 90 observations per treatment for this paper, we also collected 30 additional observations for these treatments as well as 90 new observations for the After treatment. Results remain essentially unchanged if we consider only the first 60 observations in each treatment (results available from the authors).

⁶ In one group in treatment Objective, the referee did not follow the instructions and rejected both amounts sent even though this was not part of the instructions. The experimenter only realized this at the end of the session. We decided to discard this observation. To reach the sample size we had originally planned, we ran an additional session.

via a second set of on-screen instructions that they had an opportunity to send up to \$5 of their show-up fee to the referee.

In the Before treatment, workers were informed they could put the money for the referee in a single envelope (labeled with their participant ID) together with the printed copy of their joke. In the After treatment, workers were asked to put the money and the jokes into two separate envelopes. In both cases, workers were also informed that the referee would keep their money only if they won the prize and that it would be returned to them otherwise.

After recording the monetary content of each envelope in private, the experimenter delivered the envelopes with jokes to the referees. In the Before treatment, the envelopes also contained the money sent by workers; in the After treatment, the referee received the envelopes with the money two minutes after the envelopes with the joke.

In the Objective treatment, participants had five minutes to identify the color of as many words as possible using the computer keyboard. We showed participants a sequence of color words on screen (e.g., blue, red, yellow) one after the other and asked them to identify the printed color of each word as quickly as possible. We used a congruent version of the task, meaning that the color word and its printed color were compatible (e.g., blue was always written in blue letters). The number of correctly identified words determined the worker's score for the task. The worker's final score was printed on a score sheet using a scatter plot, where a dot on a random coordinate in the plot represented each correctly identified word. The workers' instructions regarding the money were the same as in the Before treatment. In particular, workers had to put the printed scatter plot and the money in one envelope that would be delivered to the referee.

In all three treatments, the instructions informed the referees that they could only keep the winners' money and had to return the losers' money by putting it back into the loser's envelope. The referees had five minutes to determine the winner, after

which all envelopes were returned to the experimenter who then recorded their decisions. In the Before and After treatments, we also asked referees to rate the quality of each joke on a scale from 0 to 10; these ratings were collected at the end of the experiment.

The experiment consisted of two rounds with the same matching of participants. To prevent referees from letting the highest amount of money sent win in round 1 for strategic reasons, no feedback was provided between rounds. Workers started the second round while the referees were evaluating their first round. The procedure for round 2 was identical to that of round 1, apart from the topic of the joke (“Psychologists”).

We subsequently recruited additional participants as independent raters. These participants had not previously participated in the experiment and were asked to rate the jokes in exchange for class credit. Each rater was presented with up to six randomly selected pairs of jokes that had “competed” in the experiment, and was asked to rate their quality on a scale of 0 to 10 and determine which was the best joke. Between 18 and 28 different raters rated each joke. This gives us an unbiased measure of quality for the Before and After treatments.

A.3. Results

In this section, we focus on the analysis of referee behavior below, using one referee as one independent observation. No significant differences in worker behavior exist across treatments, allowing us to focus on referees. In particular, there is no significant difference in the average amount of money sent across treatments (Mann-Whitney, $p > .15$) or in the distributions of amount sent (Kolmogorov-Smirnov, $p > .45$) both when we look at one of the rounds individually

or when we combine them. Furthermore, the quality of the jokes was similar in the Before and After treatments (Mann-Whitney $p > .55$; Kolmogorov-Smirnov, $p > .75$).

We use both parametric and non-parametric tests to investigate differences between treatments. For non-parametric tests involving data from both rounds, we take the average over both rounds as the unit of observation.

Joke Quality. For the non-parametric tests discussed below, we examine whether the joke with the highest quality won. For this purpose, we do not include all joke pairs because in some cases, the jokes were simply too close in quality to be reliably distinguishable. Hence we only consider two jokes within a joke pair to be sufficiently different from each other if the fraction of independent raters choosing one joke over the other as winner is different from chance at the 10% level in a test of proportions. For our minimum number of raters per pair (18), this implies taking only those pairs in which at least 65.1% of independent raters picked one of the jokes as the winner (test of proportions, $Z=1.281$, $p=0.100$).^{7,8} By this criterion, 66% of pairs over the two joke treatments combined are sufficiently different from each other. Furthermore, to facilitate direct comparisons across treatments, we also use a threshold value for the Objective treatment to exclude the performance levels that were very similar. We picked the threshold value to be 11 points, because this value includes the same fraction of data points included in the subjective treatments.

In the regression analysis that follows after the non-parametric tests, we do not use thresholds and incorporate all observations, including those in which quality was similar across the two workers.

⁷ For a threshold of 69.4%, which corresponds to jokes being significantly different at the 5% level, the results are similar. To keep the largest number of observations, we chose to focus on the threshold of 65.1% instead.

⁸ The agreement of raters is also reflected in the difference in our measure of quality, the average rating provided by the independent raters. The average difference in ratings within pairs of jokes that exceed the 65.1% threshold (1.63) is significantly larger than the average difference in quality in jokes below the threshold (0.75) (Mann-Whitney test, $p < 0.001$).

Referee Choices. Figure A.1 displays the fraction of referees choosing the worker who sent the highest amount of money (Amount Sent, left section of Figure A.1) and the fraction of referees choosing the worker with the highest quality (right section) as winner across the three treatments.

In the Before treatment, in which the incentives and the joke were received simultaneously, 84% of the workers who sent the highest amount of money won the prize, which is significantly greater than chance (Wilcoxon signed-rank (WSR) test, $p=.001$). By contrast, only 56% of the best jokes won the prize, a fraction that is not significantly different from chance (WSR test, $p=0.491$). Thus, incentives appear to distort judgment in this treatment.

In the After treatment, the percentage of workers with the highest amount sent who win decreases to 73%, which is still significantly larger than chance (Mann-Whitney (MW) test, $p=.003$) and not statistically different from the Before treatment (MW test, $p=.369$). However, the percentage of workers with the best joke who won in this treatment is 81%, which is significantly higher than what we observed in the Before treatment (MW test, $p=.027$). Further, it is significantly different from chance (WSR test, $p<.001$). This finding is consistent with our hypothesis that making motivated self-deception more difficult increases the importance of quality. The treatment difference in the importance of quality is strong and economically significant. The best joke winning 81% of the time is equivalent to 62% of referees going for quality; the corresponding percentage for the Before treatment is 12%.

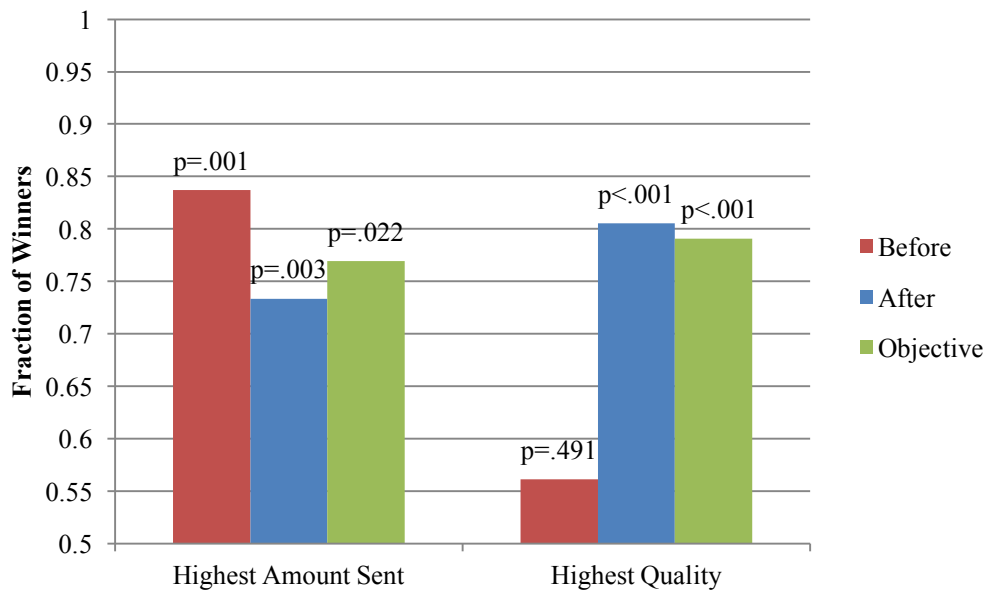


FIGURE A.1. FRACTION OF WINNERS CONDITIONAL ON HIGHEST AMOUNT SENT OR QUALITY

Notes: The p-values are calculated using a Wilcoxon signed rank test that tests if the reported fraction is significantly larger than .5. Workers are classified as having a better rating when at least 65.1% of independent raters agree their joke is better (treatments Before and After) or when their performance on the Stroop task is at least 11 words better (treatment Objective).

In the Objective treatment, 77% of workers who sent the highest amount of money won the prize (MW test, $p=.021$), which is not significantly different from either the Before (MW test, $p=.730$) or After (MW test, $p=.572$) treatments. Further, 79% of the workers with the best score on the task won (MW test, $p<.001$). This percentage is larger than the one observed in the Before treatment (56%, MW test, $p=.035$) but similar to the percentage observed in the After treatment (81%, MW test, $p=.790$). Thus, making self-deception more difficult by using an objective task also increases the importance of quality in determining the winner.

To investigate the effect of incentives and quality simultaneously, we also report the results of probit regression analyses in Table A.1. To facilitate comparisons between coefficients and treatments, we report marginal effects and have standardized all independent variables, so that the coefficients represent the effect

of a one-standard-deviation increase in the independent variable. We also allow the importance of quality to differ depending on whether or not the referee received two identical amounts of money. Intuitively, when the amounts sent are identical, referees no longer have a monetary incentive to distort the outcome and can therefore be expected to be more interested in quality.⁹

Column 1 shows that in the Before treatment, relative to a situation with equal amounts of money, increasing the difference in the amount sent by one standard deviation increases the likelihood of winning by 42 percentage points. By contrast, having the best joke does not increase the likelihood of winning when referees receive different amounts of money from the workers. However, the quality of the joke does matter when referees receive two identical amounts. This result shows that when referees no longer have an incentive to distort the outcome, they choose the better joke as the winner, whereas when incentives are in place, their judgment is biased.

Column 2 shows that the observed pattern is different in the After treatment. In contrast to the Before treatment, the quality of the joke matters even when the two amounts sent are different. Conversely, incentives matter less than in the Before treatment. Column 3 reports the results for the Objective treatment. In this treatment, when the amounts sent differ, both quality and incentives matter, with the (normalized) marginal effect for quality being somewhat larger than the coefficient for incentives. Quality also matters when the amounts of money are identical. Additional analyses are provided in section A.4 where we provide several robustness checks.

We also examine referee behavior distinguishing between cases in which referees received two positive amounts of money and one positive amount, respectively.

⁹ Because the two workers in each pair are the exact inverse observation of one another and therefore not independent observations, we randomly select one worker per pair to include in the analysis. In section A.4 below, we redo the analysis with 1,000 random samples to show that the results reported here are not due to the particular random sample that was selected.

Intuitively, justifying letting the worst performer win when both workers send money might be easier, and as a result, self-image costs might be higher when only one worker sends money. The analysis provides some support for this conjecture, as shown in section A.4.

TABLE A.1.—PROBIT REGRESSIONS FOR REFEREES

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent differ)	-.001 (.103)	.148** (.066)	.369*** (.110)
Quality Difference (amounts sent identical)	.401*** (.144)	.586*** (.186)	.229** (.116)
Amount Sent Difference	.422*** (.102)	.206** (.081)	.197** (.091)
Treatment	Before	After	Objective
Standard Errors	Clustered	Clustered	Clustered
Observations	60	60	62
Clusters	30	30	31

Notes: Probit estimates (marginal effects). Quality Difference is the difference between the quality of the joke (i.e., the average score among independent raters) of the selected worker and the other worker in the group. Amount Sent Difference is the difference between the amount of money sent by the selected worker and the amount sent by the other worker in the group. In each specification, we randomly select one worker per referee per round. Robust standard errors are clustered at the referee level.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level.

Quality Ratings. As mentioned above, in addition to asking referees to determine the winning worker, we also asked referees in the Before and After treatments to rate the quality of both jokes on a scale from 0 to 10. This measure was not incentivized. Interestingly, the correlation between the referees' ratings and the grades given by independent raters is 0.27 for the Before treatment and 0.54 for the After treatment. An OLS regression with ratings from the referees as a dependent variable and ratings from the independent raters as an independent variable shows this correlation is much stronger for the After treatment ($\beta=1.19$, $p<.001$) than for the Before treatment ($\beta=.46$, $p=.019$); including an interaction term between treatment and independent ratings shows that the difference in coefficients is

significant ($\beta=.73$, $p=.006$). Thus, referees in the After treatment gave a less biased judgment of joke quality than referees in the Before treatment. This finding is in line with self-deception being harder in the After treatment as well.

Taken together, these results are in line with motivated self-deception. As in the Distorted Advice Experiment, when the task is subjective, quality plays a larger role in determining a winner when referees perform their judgments before being aware of the incentives. When incentives are provided at the same time as jokes, referees' judgment shifts toward workers who sent the highest amount of money. Conditional on amounts sent being different, quality no longer plays a role. Additionally, when the task is more objective, receiving the incentives together with the task does not lead to the same bias.

A.4. Robustness checks

We investigate differences in the effect of quality of jokes and the effect of receiving money on referees' choices across treatments, using OLS regressions and interacting these variables with treatment dummies. In this analysis, we use OLS rather than probit to facilitate treatment comparisons. Table A.2 reports the results. Column 1 suggests that the difference in amount sent is a less important determinant of referees' choices in the After treatment than in the Before treatment ($p=.11$).¹⁰ Conversely, quality difference between jokes plays a larger role in the After treatment than in the Before treatment ($p=.072$). Column 2 shows that a similar pattern emerges when comparing the Before treatment with the Objective treatment: the difference in amount sent by the two workers is more important in the Before treatment ($p=.047$), whereas quality plays a larger role in the Objective

¹⁰ The significance of this coefficient varies depending on the random sample drawn. In 575 out of 1,000 random samples, the coefficient is significant at the 10% level or lower. In the draw randomly selected for Table A.2, the coefficient is not significant. All other interaction terms are robust and remain significant in at least 900 random samples out of 1,000.

treatment. Finally, column 3 shows that amount sent and quality have similar effects in the Objective and After treatments.

TABLE A.2—OLS INTERACTION TERMS FOR REFEREES

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent differ)	-.019 (.058)	-.019 (.058)	.219*** (.060)
Quality Difference (amounts sent identical)	.244*** (.048)	.244*** (.048)	.222** (.090)
Amount Sent Difference	.298*** (.039)	.298*** (.039)	.136* (.069)
Quality Diff. (amounts sent differ) * After	.152* (.079)		-.084 (.086)
Quality Diff. (amounts sent identical) * After	.155* (.085)		.174* (.110)
Amount Sent Difference * After	-.122 (.076)		.040 (.095)
Quality Diff. (amounts sent differ) * Objective		.239*** (.083)	
Quality Diff. (amounts sent identical) * Objective		-.022 (.101)	
Amount Sent Difference * Objective		-.162** (.080)	
Treatment	Before & After	Before & Objective	Objective & After
Standard Errors	Clustered	Clustered	Clustered
Observations	120	122	122
Clusters	60	61	61

Notes: OLS estimates. Quality Difference is the difference between the quality of the joke (i.e., the average score among independent raters) of the selected worker and the other worker in the group. We standardize this variable to have the same mean and standard deviation in the objective task as in the joke task. Amount Sent Difference is the difference between the amount of money the selected worker sent and the one sent by the other worker in the group. In each specification, we randomly select one worker per referee per round. The regressions also include treatment dummies to correct for difference in the overall fraction of winners as the result of randomly selecting workers; their coefficients are always small and not significant. Robust standard errors are clustered at the referee level.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

For the analyses reported in Table A.1, we randomly selected one worker for each pair. To ensure that our results are not due to the specific random sample selected for the analysis, we additionally redo the regressions reported in Table A.1 with 1,000 different random samples and report the average results as well as the standard deviation in the estimated marginal effects (Table A.3). The results are very similar to those reported in Table A.1, which reveals that our results are robust to the particular random sample we used.

TABLE A.3 —PROBIT REGRESSIONS FOR ALTERNATIVE RANDOM SAMPLES

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent differ)	.00 [.01]	.15 [.01]	.36 [.02]
Quality Difference (amounts sent identical)	.41 [.03]	.56 [.04]	.23 [.02]
Amount Sent Difference	.43 [.02]	.20 [.02]	.20 [.01]
Treatment	Before	After	Objective
Observations	60	60	62
Clusters	30	30	31

Notes: Probit estimates (marginal effects). Quality Difference is the difference between the quality of the joke (i.e., the average score among independent raters) of the selected worker and the joke of the other worker in the group. Amount Sent Difference is the difference between the amount of money sent by the selected worker and the one sent by the other worker in the group proposed. In each specification, we re-estimate the regression reported in Table A.1 1,000 times with different random samples of one worker per referee per round; the marginal effect is the average of the 1,000 marginal effect estimates, and the number in square brackets is the standard deviation of the 1,000 marginal effect estimates.

So far, our analysis has looked only at differences in the monetary amount sent and quality (in regressions) or the effect of one amount/performance being greater than the other (in non-parametric tests). However, referees might respond differently to quality when they receive two strictly positive monetary amounts compared to situations in which only one person sends a positive amount. For example, a referee might be happy to take \$5 over \$2 when both workers send money, but not \$3 over \$0, because the dishonesty of this act is more salient. Although justifying taking the highest amount sent is relatively easy when both workers send a positive amount, because they are both being dishonest, justifying taking the highest amount sent when one of the participants behaves honestly may be harder.¹¹

In Table A.4, we estimate separate coefficients for both the amount sent difference and quality difference for the cases when two workers or only one

¹¹ We did not originally intend to incorporate this analysis in our paper; we only included it after it had been repeatedly suggested to us by seminar participants and others.

worker sent a positive amount, respectively. In the Before treatment, differences in amount sent are always important and quality only matters when the amounts sent by two workers are identical. In the After treatment, a shift occurs in the relative importance of quality and amount sent. It is especially strong for the cases in which only one worker sent a positive amount of money. In such cases, quality matters and the amount sent to the referee does not. By contrast, when both workers send positive amounts, the effect of quality is not statistically significant and the effect of amount sent is.

Figure A.2 illustrates this result graphically. When the referee receives a positive amount from one worker only, the better-quality joke wins 63% of the time in the After treatment (WSR test, $p=.248$), compared to 93% in the Before treatment (WSR test, $p=.004$). This difference is statistically significant (MW test, $p=.049$). By contrast, when referees receive positive amounts by both workers, the better joke does not win significantly more (or less) often than chance in either treatment.

For the Objective treatment, the effect of amount sent is similar to the After treatment: it matters only when two workers send positive amounts. However, in this case, the importance of quality does not seem to depend on whether one worker or two workers sent a positive amount.

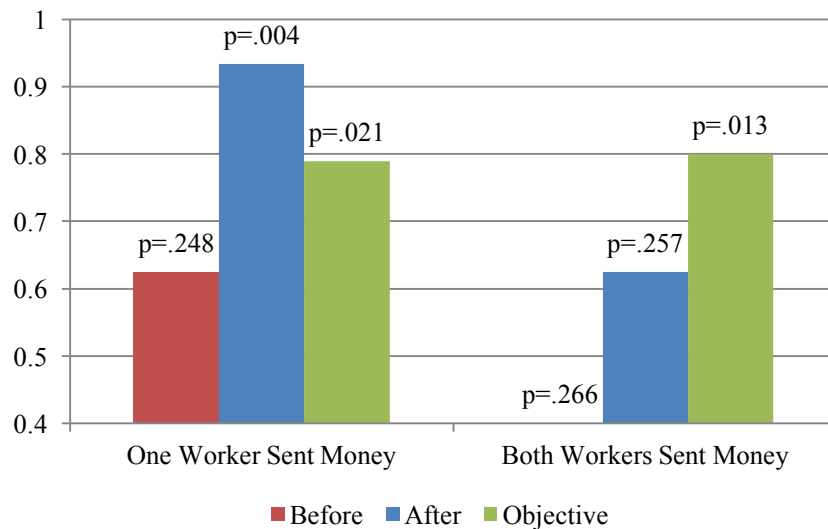


FIGURE A.2. FRACTION OF WINNERS CONDITIONAL ON ONE WORKER VS. BOTH WORKERS SENDING MONEY.

Notes: The p-values are calculated using a Wilcoxon signed rank test that tests if the reported fraction is significantly larger than .5. Workers are classified as having a better rating when at least 65.1% of independent raters agree their joke is better (Before and After treatments) or when their performance on the Stroop task is at least 11 words better (treatment Objective).

TABLE A.4—PROBIT REGRESSIONS ONE VERSUS TWO POSITIVE AMOUNTS SENT

Probability (winning)	(1)	(2)	(3)
Quality Difference (amounts sent differ)	.059	.392**	.343***
(one worker sent a positive amount)	(.169)	(.188)	(.125)
Quality Difference (amounts sent differ)	-.008	.085	.636***
(both workers sent positive amounts)	(.112)	(.113)	(.177)
Quality Difference (amounts sent identical)	.402***	.584***	.229**
	(.143)	(.187)	(.116)
Amount Sent Difference	.527***	.140	.177
(one worker sent a positive amount)	(.156)	(.113)	(.124)
Amount Sent Difference	.349***	.465***	.513**
(both workers sent a positive amount)	(.136)	(.156)	(.226)
Treatment	Before	After	Objective
Selected Workers	Random	Random	Random
Standard Errors	Clustered	Clustered	Clustered
Observations	60	60	62
Clusters	30	30	31

Notes: Probit estimates (marginal effects). Quality Difference is the difference between the average quality of the joke (as judged by the independent raters) of the selected worker and that of the other worker in the group. Amount Sent Difference is the difference between the amount of money sent by the selected worker and the amount of money sent by the other worker in the group. All marginal effects are evaluated at the means for all independent variables. Robust standard errors are clustered at the referee level.

*** Significant at the 1% level; ** Significant at the 5% level; * Significant at the 10% level

Appendix B: Instructions

In this section, we present the instructions for the experiments. We first present the instructions for the Distorted Advice experiment, separately for advisors and clients. The instructions for the Strict Dominance and Weak Dominance Experiment are the same except for the change in the payoffs of Product B. We then present the instructions for the additional experiment presented in Appendix A, again separately for workers and referees. Finally, we also present the instructions for the independent raters who rated the jokes of the additional experiment.

B.1. Advisor – Client Game

The instructions for advisors are presented in Subsection B.1.1. Treatment variations are added in squared brackets, indicating the treatment they apply to. The instructions for clients are presented in Subsection B.1.2.

B.1.1. Instructions for Advisors

Decision-making study: Description

Your role

In this study, you have been assigned the role of advisor. You will be asked to recommend one of two products to another participant, the client, who will participate in another session of this study.

How it works

- As advisor you are informed about the details of two products, product A and B
- You recommend one of the products, A or B, to the client

- The client, who does not have any information about A and B, receives your recommendation
- The client chooses product A or B. The client's choice and chance will determine his/her payoffs.

Pairing between advisors and clients

At the end of this study, we will randomly select one advisor out of ten and give his or her recommendation to a client, who will be paid accordingly. As mentioned above, the client will be a participant in another session of this study.

Payments

You will receive \$1 for providing your recommendation.

[Before: If you recommend product A, you will receive an additional \$1 as commission.]

---next page---

Information about the two products

Product A

- Product A is produced by Company A and has two potential qualities, low or high
- If the quality of product A is low, it will yield \$2 to the client.
- However if the quality of product A is high, it will yield \$4.
- There is an equal chance that the quality will be either low or high.
- Whether the quality of the product is actually high or low will be determined by a coin toss later.

Product B

- Product B is produced by Company B
- As product A, it has two potential qualities, low or high.
- If the quality of product B is low, it will yield \$1 to the client.

- However if the quality of product B is high, it will yield \$7.
- There is an equal chance that the quality will be either low or high.
- Whether the quality of the product is actually high or low will be determined by a coin toss later.

Hence, after the client chooses between product A or B, we will use the outcome of a coin toss, made by a volunteer, to determine his/her payoff.

Please consider your recommendation

In a few minutes you will be given a decision sheet and you will be asked to complete the sentence:

“I recommend you to choose product (A or B) _____”

This decision sheet will be shown to the client before he or she chooses between product A or B. You will be asked to put it in an envelope and the envelope will be delivered to the client.

Please take a minute to decide which product to recommend. Click the arrow below when you are ready to provide your recommendation.

---next page---

Please raise your hand now

The experimenter will give you your decision sheet where you can write your recommendation to the client.

Once you have your decision sheet, click below to proceed to the next screen.

---next page---

Decision sheet

Please write down your recommendation on the decision sheet

[Before and After: **If you recommend product A, you will receive an additional \$1 as a commission.**]

Your recommendation

To make sure all records are kept, please input your recommendation on this screen as well.

“I recommend you to choose product

B.1. 2. Instructions for Clients

Decision-making study: Description

Your role

Welcome to this study on decision-making. In this experiment you are matched with another participant. Neither your identity nor the identity of the participant you are matched with will be revealed.

In this study, you have been assigned the role of the client. Your task will be to choose one of two products, which will result in some monetary payments to you. The monetary payment you will receive depends on the product you choose.

In a previous session of the study another participant, the advisor, was provided with information about the two products and was asked to recommend a product to you.

In a moment, you will receive the recommendation from the advisor. Please raise your hand.

----- After receiving the recommendation -----

Which product do you choose?

Product A

Product B

B.2 Distorted Choice Game

The instructions for the Before and After treatment are presented in subsections B.2.1 (instructions for workers) and subsections B.2.2 (instructions for referees). The instructions for treatment Objective are presented in subsections B.2.3 and B.2.4. Finally, the instructions for independent raters are presented in subsection B.2.5.

B.2.1 Treatment Before and After: Instructions for Workers

Welcome to today's experiment. Please read the instructions carefully. If you have any questions, please raise your hand and one of the experimenters will come to your desk to answer your question.

----- next screen -----

You have been assigned to the role of Participant A. For the remainder of the experiment you will be matched with two other participants: Participant B and a Referee. The Referee will now be moved to a different room.

----- next screen -----

On your desk you can find an envelope with 10 dollars. This is your show-up fee for taking part in this experiment. Both you and Participant B have received a \$10 show-up fee whereas the Referee has received a \$5 show-up fee. Please do not remove the money from the envelope until you are instructed to.

Both you and Participant B will be asked to work on a task for two rounds. The task will be explained below. After each round the Referee will decide whether you or Participant B performed the task better. The Participant that performed better (as decided by the Referee) will receive an additional \$10 prize on top of the show-up fee. The other Participant will receive nothing.

You will be matched to the same Referee and Participant B in both rounds. None of you will ever know the identity of the other two participants.

Do you have any questions before we explain the task to you?

----- next screen -----

Your task:

Your task is to come up with a joke about a certain topic, which will be announced after the instructions. In total, you will have 10 minutes to come up with a joke. The joke can be short or long, a simple one liner or a full anecdote. The experimenter will let you know when you have 5 minutes as well as 1 minute left for the round.

----- next screen -----

Check-up questions

How much will you earn (in dollars) in a given round if you are the winner?

Who is going to evaluate your task?

True or false: the Referee and Participant B will be the same participants in both rounds of the experiment.

----- next screen -----

You are now ready to start the experiment. Please raise your hand when you are ready to start the task. Do not proceed to the next page. The experimenter will instruct you to start when the other participants have finished reading the instructions.

----- next screen -----

Please write a joke about economists. You have 10 minutes to complete the task.

----- next screen -----

What do you believe is the probability that you wrote better jokes than Participant B?

----- next screen -----

Please wait while we are printing your joke.

After you have received your joke, please put it into the large envelope with the number so it can be handed over to your Referee. You also have the option to add money for the Referee [*Before*: and put it in the envelope with the joke] [*After*: and put it into the small envelope on the right side of your desk]. For this purpose, you can take up to \$5 out of the smaller envelope with your show-up fee [*Before*: and put it into the larger envelope together with the joke you wrote] [*After*: and put it into the other small white envelope that is on your desk]. Participant B also has the option to add up to \$5 to the envelope he/she sends to the Referee.

[*Before*: The Referee will be given both your envelope with the joke and the money and Participants B's]. [*After*: The Referee will be given both your and Participant B's large envelope with the joke and the small envelope with money.]. He/she will be asked to read the jokes and decide which one wins. If the Referee chooses your joke, then you will get an additional \$10 and the Referee will keep the money you sent him/her. Participant B will get the money he/she sent to the Referee back. If the Referee chooses Participants B's joke, then Participant B will get an additional \$10 and the Referee will keep the money he/she sent to him. In this case you will get back the money that you sent to the Referee.

Please raise your hand when the [*Before*: envelope for the Referee is ready.] [*After*: envelopes for the Referee are ready]. The experimenter will bring it to the Referee in the next room. After the Referee has determined the winner, the envelope will be collected by the experimenter. The envelope will be returned to you after the Referee has finished grading the second round of jokes.¹²

¹² Instructions for round 2 started from "please write a joke about economists" onwards and were identical to the instructions for round 1, except that workers were instructed to write a joke about psychologists instead. After round 2, they were asked to fill out a short demographic questionnaire, wait for the referees to finish determining the winners and then collect their payment.

B.2.2. Treatment Before and After: Instructions for Referees

Welcome to today's experiment. Please read the instructions carefully. If you have any questions, please raise your hand and one of the experimenters will come to your desk to answer your question.

----- next screen -----

You have been assigned to the role of the Referee. For the remainder of the experiment you will be matched with two other participants: Participant A and Participant B.

Please raise your hand. The experimenter will escort you to a different room.¹³

----- Paper instructions: page 1 -----

Welcome to the experiment. On your desk you will find a small envelope with \$5. This is your show-up fee for taking part in this experiment; Participant A and Participant B have received a \$10 show-up fee for the experiment.

Today your task is to rate the quality of a joke written by Participant A and a joke written by Participant B. You will be matched to the same Participant A and Participant B in both rounds. None of you will ever know the identity of the other two participants.

Both Participants have 10 minutes to write a joke. After Participants A and B have finished their jokes, they will print them and put them in an envelope which will be brought to you by an experimenter.

You will then have 5 minutes to read both jokes and determine who of the two did the best job, i.e. determine the winner. The winner will receive a prize of \$10, whereas the loser will receive nothing. Please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope.

¹³ After this part, referees were taken to a different room where the remainder of their instructions was provided on separate sheets of paper.

You will also be asked to rate the quality of both the winner's and the loser's joke on a scale from 0 to 10 (on the evaluation form).

The envelopes will then be collected by the experimenter and you will be asked to grade a second round of jokes, written by Participants A and B while you were grading.

The envelopes for both rounds will be returned to Participants A and B at the end of the second round.

The topic for the first round will be 'economists', the topic of the second round will be announced to you after you finish grading the first round.

Please wait while Participants A and B finish writing their jokes. If you have any questions in the meantime, please ask the experimenter.

----- Paper instructions: page 2¹⁴ -----

In a few moments you will receive two envelopes containing the jokes written by Participants A and B. To grade their jokes, please indicate your rating for both Participants on the evaluation form on a scale from 0 to 10.

Participants A and B also have the opportunity to [*Before*: add money to their envelope] [*After*: send you money in an envelope]. You have the option to keep the money sent to you by either Participant A or Participant B. If you keep the money of a Participant, he or she will automatically be the winner. The loser's money will then be returned.

After determining the winner, please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope. After five minutes, an experimenter will collect the envelopes. The envelopes will be returned to Participants A and B at the end of the second round.

Please remain patient while we are printing the jokes.¹⁵

¹⁴ This page was only delivered after participants A and B had finished typing up their joke.

¹⁵ Instructions for round 2 were provided after the experimenters had collected the first round of envelopes. The instructions contained the topic of the second round. Otherwise, they were identical to the last page of the instructions for round 1 (from "In a few moments" onwards). After referees finished grading the second round, they were given a demographic

----- Paper instructions: page 3 -----

[*After*: Please read the jokes for Participant A and B. We will bring you the envelopes with the money in few minutes.]

B.2.3. Treatment Objective: Instructions for Workers

Welcome to today's experiment. Please read the instructions carefully. If you have any questions, please raise your hand and one of the experimenters will come to your desk to answer your question.

----- next screen -----

You have been assigned to the role of Participant A. For the remainder of the experiment you will be matched with two other participants: Participant B and a Referee.

The Referee will now be moved to a different room.

----- next screen -----

On your desk you can find an envelope with 10 dollars. This is your show-up fee for taking part in this experiment. Both you and Participant B have received a \$10 show-up fee whereas the referee has received a \$5 show-up fee. Please do not remove the money from the envelope until you are instructed to.

Both you and Participant B will be asked to work on a task for two rounds. The task will be explained below. After each round the Referee will decide whether you or Participant B performed the task better. Your goal is to complete as many words as possible in 5 minutes.

The Participant that performed better (as decided by the Referee) will receive an additional \$10 prize on top of the show-up fee. The other Participant will receive

questionnaire to fill out and had to sign for payment in their individual rooms a few minutes later, after which they were allowed to leave the laboratory.

nothing. You will be matched to the same Referee and Participant B in both rounds. None of you will ever know the identity of the other two participants.

Do you have any questions before we explain the task to you?

----- next screen -----

During each round of the experiment you will be shown a sequence of words. These words will be printed in different colors: yellow, blue, purple, orange, or red. Your task is to indicate the color of each word. Only the colors named correctly will count towards your total. This task will last for a total of 5 minutes.

You can indicate the color of your choice using the keyboard. The relevant keys are y (for yellow), r (red), p (purple), o (orange) and b (blue). The key-color combinations will also be visible at the bottom of the screen throughout the task. Be aware: if you press any key other than the one corresponding to the correct color, this will not be counted as a correct response. This also holds for keys that do not refer to any color. On the next page you will have the opportunity to practice the task with a sequence of 10 words.

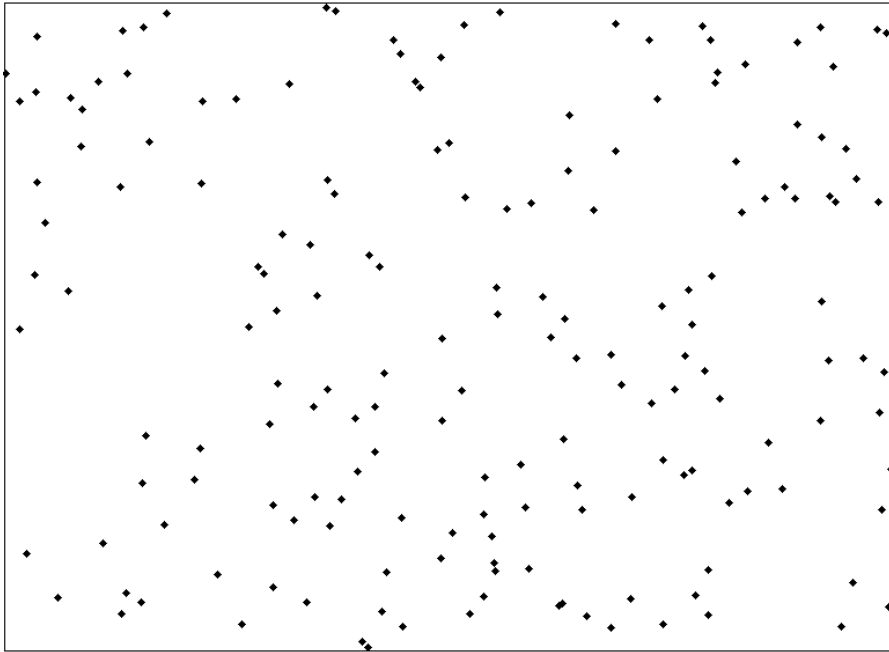
----- next screen -----

*Practice screen*¹⁶

----- next screen -----

After you finish the task, your score will be printed on a score sheet that will be handed over to your referee. Your score sheet will be similar to the example below. Every color you successfully indicated will be represented by a dot on the score sheet.

¹⁶ Here, workers had the opportunity to practice the task with 10 words. They could only move forward once they had gone through all 10 words.



----- next screen -----

Please answer the following questions before proceeding to the next page.

Question 1: How much will you earn (in dollars) in a given round if you are the winner?

Question 2: Who is going to evaluate your task? Participant A, Participant B or the Referee?

Question 3: True or false: the Referee and Participant will be the same participants in both rounds of the experiment.

----- next screen -----

You are now ready to start the experiment.

Please raise your hand when you are ready to start the task.

Do not proceed to the next page. The experimenter will instruct you to start when the other participants have finished reading the instructions.

----- next screen -----

Task –see a sample screenshot below



----- next screen -----

What do you believe is the probability that you have a better score than Participant B?

----- next screen -----

Please wait while we are printing your score sheet.

After you have received your score sheet, please put it into the large envelope with the number so it can be handed over to your Referee. You also have the option to add money for the Referee and put it in the envelope with the score sheet. For this purpose, you can take up to \$5 out of the smaller envelope with your show-up fee and put it into the larger envelope together with your score sheet. Participant B also has the option to add up to \$5 to the envelope he/she sends to the Referee.

The Referee will be given both your envelope with the score sheet and the money and Participant B's envelope. He/she will then be asked to determine which Participant wins. If the Referee decides that you win, then you will get an additional \$10 and the Referee will keep the money you sent him/her. Participant B will get back the money he/she sent to the Referee. If the Referee decides that Participant B wins, then Participant B will get an additional \$10 and the Referee will keep the

money he/she sent to him. In this case you will get back the money that you sent out to the Referee.

Please raise your hand when the envelope for the Referee is ready. The experimenter will bring it to the Referee in the next room.

----- next screen -----

You are now ready to start round 2. This round will be similar to round 1: you will again have to indicate the color of a sequence of words and the task will again be graded by the Referee. Please remember that you will be matched to the same Referee and the same Participant as before. You will again have the option to send money to the referee after you finish your task.¹⁷

B.2.4. Treatment Objective: Instructions for Referees

Welcome to today's experiment. Please read the instructions carefully. If you have any questions, please raise your hand and one of the experimenters will come to your desk to answer your question.

----- next screen -----

You have been assigned to the role of the Referee. For the remainder of the experiment you will be matched with two other participants: Participant A and Participant B.

Please raise your hand. The experimenter will escort you to a different room.

----- Paper instructions: page 1 -----

Welcome to the experiment. On your desk you will find a small envelope with \$5. This is your show-up fee for taking part in this experiment; Participant A and Participant B have received a \$10 show-up fee for the experiment.

¹⁷ Instructions for round 2 were the same as round 1, starting from "what do you believe is the probability ...". After round 2, they were asked to fill out a short demographic questionnaire, wait for the referees to determine the winners and then collect their payment.

Today your task is to determine the score of Participant A and Participant B on a task. You will be matched to the same Participant A and Participant B in both rounds. None of you will ever know the identity of the other two participants.

Participant A and B's task is to determine the color of a series of words displayed on their computer screen. The participants will be shown a sequence of words one at the time and they will have to indicate the colors of the words. Their goal is to complete as many words as possible in 5 minutes. A screenshot of the task has been provided to you on a separate sheet.

Both Participants have 5 minutes for the task. After 5 minutes, their scores will be printed on a score sheet and each one of them will get his or her own printout. The printout score sheet will be similar to the sample score sheet provided to you as an example. Each color successfully determined by the participants will be represented by a single dot on the score sheet. Each participant will then put his/her score sheet in an envelope that will be brought to you by the experimenter.

You will then have 5 minutes to determine the winner. The winner will receive a prize of \$10, whereas the loser will receive nothing. Please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope.

The envelopes will then be returned to Participants A and B and you will be asked to grade a second round of score sheets representing the number of colors successfully indicated by Participants A and B while you were grading.

Please wait while Participants A and B complete the first round. If you have any questions in the meantime, please ask the experimenter.

----- Paper instructions: page 2 -----

In a few moments you will receive two envelopes containing the score sheets of Participants A and B.

Participants A and B also have the opportunity to add money to their envelope. You have the option to keep the money sent to you by either Participant A or Participant B. If you keep the money of a Participant, he or she will automatically be the winner. The loser's money will then be returned.

After determining the winner, please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope. After five minutes, the experimenter will collect the envelopes and return them to Participants A and B in the other room.

Please remain patient while we are printing the score sheets.

----- Paper instructions: page 3 -----

Please wait while Participants A and B are finishing the second round. After Participants A and B have finished the second round, the procedure will be similar to round 1.

You will again receive two envelopes containing the score sheets of Participants A and B.

Participants A and B also again have the opportunity to add money to their envelope. You have the option to keep the money sent to you by either Participant A or Participant B. If you keep the money of a Participant, he or she will automatically be the winner. The loser's money will then be returned.

After determining the winner, please make sure to indicate the winner by placing a winner card in the winner's envelope and a loser card in the loser's envelope. After five minutes, an experimenter will collect the envelopes and return them to Participants A and B in the other room.

Please remain patient while Participants A and B are finishing the second round.¹⁸

¹⁸ After referees finished grading the second round, they were given a demographic questionnaire to fill out and had to sign for payment in their individual rooms a few minutes later, after which they were allowed to leave the laboratory.

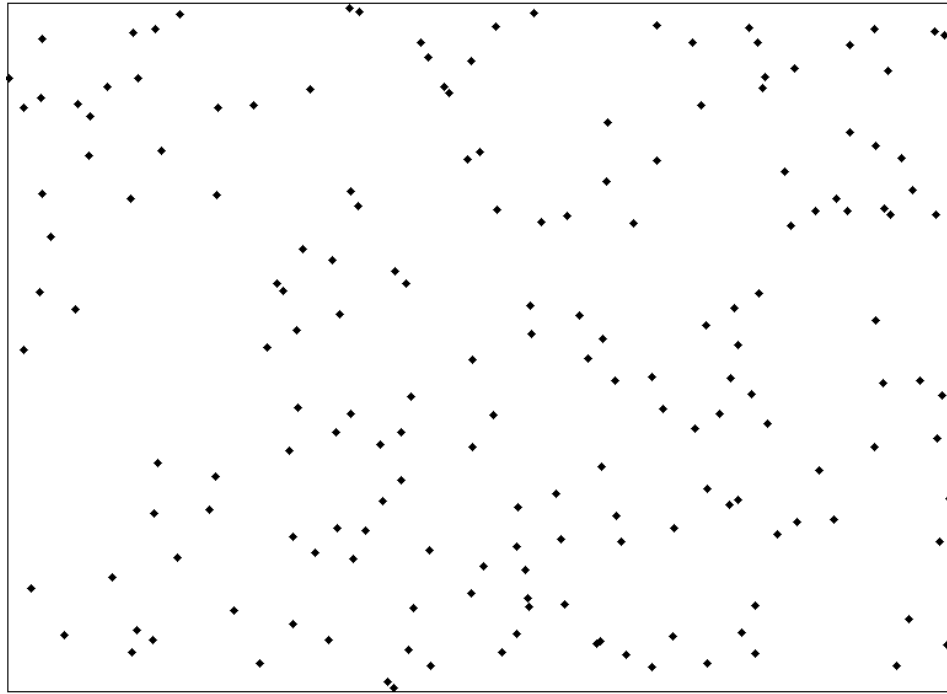


FIGURE B1. SAMPLE SCORE SHEET FOR TREATMENT OBJECTIVE

Notes: the figure was used as an example for participants in the experiment. It was generated using the same procedure as the score sheets used in the experiment, with every dot representing one correct response on the task. Referees received the score sheet on a separate page with their initial instructions.

B.2.5. Instructions for Independent Raters

Welcome!

In this experiment you will be shown six pairs of jokes. Jokes in each pair will either be about economists or about psychologists. Participants in a previous experiment wrote the jokes in 10 minutes. For each pair of jokes, you will be asked to rate the quality of both jokes and to indicate which one is better.

----- next screen -----

Please rate the quality of the following jokes about economists (psychologists).
Make sure to read both jokes before rating.

What is the quality of this joke? (0-10)

Which one is the best joke? (Joke A, Joke B)

Appendix C: Examples of Jokes

In this section, we present 9 examples of jokes written by participants in the experiment. The following jokes are the three jokes with the lowest rating, the three jokes with the highest rating, and three median jokes (as determined by the independent raters) respectively. All other jokes are available upon request.

C1. Bad Jokes

(rating: 0.83/10)¹⁹

A group of economists were having a discussion about (rating: 0.91/10)

One economist one day went to the shopping centre to buy a keyboard, the price labbed on the hat was \$59.99. While the keyboard is using solar as its battery, he start to computer the profit he can get from the keyboard. Since the solar keyboard is much expensive than the normal one, he think that he can use it 3 years, and if he uses the normal keyboard the battery is ... As he thinking, here is a college student came to the store, he bought the keyboard without thinking, and the solar keyboard is out of stock! (rating: 1.11/10)

C2. Median Jokes

economists are so fat, when they sit on a rainbow, the skittles come out
(rating: 3.5/10)

Bob, Joe, John are best friends. They all are successful to some extent. Joe is an engineer, John is a brain surgeon and Bob is an economist. Joe and John always make fun of Bob, but bob is happier than both of them.

Why is bob happy?

In college, economy majors had the most sex, studies show a whopping 4.9/5.
(rating: 3.5/10)

¹⁹ This participant did not write a joke. This was the only joke in the experiment that was empty.

3 STUDENTS ARE ASKED TO COME INTO A COMPUTER LAB TO WRITE A JOKE. THE FIRST STUDENT IS A ECON MAJOR, THE SECOND IS A BIOLOGY MAJOR, AND THE THIRD IS A PSYCH MAJOR. AFTER 10 MINUTES OF GRUELING WORK, THE THREE STUDENTS COME OUT OF THE LAB WITH THEIR JOKES IN HAND. HERE IS WHAT THEY WROTE:
ECON STUDENT: KNOCK KNOCK, WHO'S THERE? THINKA, THINKA WHO? THINK I'M GONNA WIN! / BIOLOGY STUDENT: WHAT DID THE OCEAN SAY TO THE SAND? NOTHING, IT JUST WAVED!
PSYCH MAJOR: WHAT GAMES ARE YOU TRYING TO PLAY WITH ME?
(rating: 3.52/10)

C3. Best Jokes

I asked my professor, "What are the differences between Microeconomists and Macroeconomists?" He replied, " Well, Microeconomists are usually wrong about very specific things, whereas Macroeconomists are usually wrong about things in general." (rating: 6.54/10)

What did the economist say to the girl at the bar?

Answer: You're an opportunity cost I don't want to lose. / (rating: 6.67/10)

A psychologist was conducting a group therapy session about addictions and obsessions, in which four mothers and their children were participating. Each of the mothers was asked by the psychologist to share their obsession as well as their kid's names.

The first mother said, "I am obsessed with eating, and my daughter's name is Candy."

The second mother said, "I am obsessed with money, and my daughter's name is Penny."

The third mother said, "I am an alcoholic, my daughter's name is Brandy."

The fourth mother got up, took her son by the hand, and whispered in his ear, "Come on, Dick, let's go home." (rating: 8/10)