Chen, Zhuoqiong; Gesche, Tobias

Working Paper

Persistent bias in advice-giving

Working Paper, No. 228

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Chen, Zhuoqiong; Gesche, Tobias (2016) : Persistent bias in advice-giving, Working Paper, No. 228, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-124325

This Version is available at:
https://hdl.handle.net/10419/162431

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 228

Persistent bias in advice-giving

Zhuoqiong (Charlie) Chen and Tobias Gesche

June 2016

Persistent bias in advice-giving

Zhuoqiong (Charlie) Chen
London School of Economics*

Tobias Gesche
University of Zurich**

June 2016

–comments welcome–

Abstract

We show that a one-off incentive to bias advice has a persistent effect on advisers' own actions and their future recommendations. In an experiment, advisers obtained information about a set of three differently risky investment options to advise less informed clients. The riskiest option was designed such that it is only preferred by risk-seeking individuals. When advisers are offered a bonus for recommending this option, half of them recommend it. In contrast, in a control group without the bonus only four percent recommend it. After the bonus was removed, its effect remained: In a second recommendation for the same options but without a bonus, those advisers who had previously faced it are almost six times more likely to recommend the riskiest option compared to the control group. A similar increase is found when advisers make the same choice for themselves. To explain our results we provide a theory based on advisers trying to uphold a positive self-image of being incorruptible. Maintaining a positive self-image then forces them to be consistent in the advice they give, even if it is biased.

Keywords: advice-giving, conflict of interest, self-signaling, self-deception

JEL Classification: C91, D03, D83, G11

* z.chen16@lse.ac.uk, Department of Management, London School of Economics and Political Sciences
** tobias.gesche@econ.uzh.ch, Department of Economics, University of Zurich

Doctor, should I go for the surgery or take the more gentle but less reliable alternative therapy?

As a retirement adviser, do you recommend to increase the share of stocks in my pension plan?

Professor, should we trust in the security of nuclear plants or shut them down?

1 Introduction

When making risky decisions, we often seek advice. Doctors, investment advisers, scientists, and other experts have specific skills and knowledge to assess the potential consequences of important choices. Their job is to use their specialized information and skills to provide recommendations which are supposed to be in the best interest of patients, investors, politicians, and other clients. However, advisers may face a conflict of interest. Often, third parties pay commissions or create situations such that advisers owe them and then bias advice in their interest.[1] Advisers who give in to such third-party incentives can morally accommodate this behavior by convincing themselves that they would have given the same advice, even if there had not been such a conflict of interest. For example, when a financial adviser recommends an investment fund as opposed to a less risky asset because of a sales commissions, this can later be justified by believing that it would have been the appropriate advice anyway. However, to uphold such a justification, the adviser has to act consistently. That is, an adviser has to issue the same biased advice even when the conflict of interest does not exist anymore.

This paper presents evidence for such persistent effects from advisers' conflicts of interest. In an experiment, we offer advisers a bonus which pays if they recommend less informed clients an investment option that is preferred only by risk-seeking individuals. Among advisers in a control group without such a bonus, almost no-one recommends this risky option. In contrast, almost half of the advisers to whom the bonus was offered do recommend it. Afterwards, advisers have to choose for themselves among the same options and then make a second recommendation for another client. For these tasks, it was explicitly stated that there would not be any bonus. Our results show that advisers who were previously exposed to the bonus were six times more likely to recommend the risky option than those who were not. We also find a similar increase in the probability that advisers choose the risky option for themselves. In consequence, being exposed to a conflict of interest in advice-giving in one single instance creates an externality on the advice which another client receives and the adviser's own choices.

[1] For example, US financial advisers administered more than $38 trillion for more than 14 million clients in 2011 (SEC, 2011). Despite laws like the Dodd-Frank Act which require them to "[...]*to act in the best interest of the customer*" (United States Congress, 2010, Sec. 913g), they receive sales commissions and bias their advice accordingly (Mullainathan et al., 2012; Malmendier and Shanthikumar, 2014). Other experts face such conflicts of interest too: Although supposed to be impartial, doctors reciprocate gifts from pharmaceutical companies (Dana and Loewenstein, 2003; Cain and Detsky, 2008) and scientists are dependent on industries sponsoring their research (Hilgartner, 2000; Taylor and Giles, 2005).

We present a behavioral mechanism which can explain such persistent effects on repeated advice and advisers' own choices. It is based on the human tendency to interpret own actions to infer one's own morality (Mazar et al., 2008; Benabou and Tirole, 2011). To avoid a negative and immoral self-image, biased advisers can perceive their recommendations as those which they actually should have recommended, had they actually been impartial. However, when advisers morally accommodate their corrupted behavior in such a manner, they have to stick with their advice. The reason is that changing it, in particular when the conflict of interest disappears, would signal to themselves that their initial advice was corrupted, and therefore, that they acted immorally.

Our results also show more exactly what advisers take as a reference for giving impartial advice and thus, how they try to keep a positive self-image: In principle, an adviser can internally disguise the fact that his advice was biased by forming a motivated belief (Kunda, 1990) about the clients' preferences, for example that a client is sufficiently risk-seeking.[2] In the adviser's view it is then in the client's best interest and therefore moral to recommend the risky option, even though the actual motive is the conflict of interest. This would not put the adviser under any pressure to act accordingly for himself, since his motivated belief is only about the client's risk preferences, not his own. However, prior research has shown that when forming beliefs about others' preferences, in particular risk preferences, we do so by starting from our own (Mullen et al., 1985; Faro and Rottenstreich, 2006). The question "What would I choose if I were in the client's situation?" then also determines what an adviser should recommend. Under such a rule, advisers who want to perceive themselves as incorruptible should then also choose for themselves what they have recommended to others. Our data indicate that this is the case: Having been exposed to a bonus leads advisers to choose the risk-seeking option more often. This is in line with the recent findings of Foerster et al. (2016) and Linnainmaa et al. (2016). In a large sample, they show that financial advisers hold the same expensive, under-performing portfolios as their clients, even after having left the industry.

Related literature: Our work combines findings from self-signaling, motivated beliefs, and self-deception to obtain new insights about their implications in the context of advice-giving. It captures the fact that people assign informational value to their actions to infer about their personal traits (Bodner and Prelec, 2003; Benabou and Tirole, 2004) and in particular their moral values (Benabou and Tirole, 2011). Self-signaling then means that actions are also influenced by the consequences they subsequently have on peoples' self-image. For example, Mazar et al. (2008) argue that we often do not lie as much as we, in principle, could because strong, outright

[2]Without referring to any actual gender roles we will call advisers and clients "he" and "she", respectively.

lies would damage our self-perception of being honest and moral persons. Gneezy et al. (2012) present the seemingly paradoxical finding that sales under a pay-as-you-want scheme are lower than under a low, fixed price. They explain the consumers' reluctance to set a sufficiently low pay-as-you-want-price with consumers' desire to not perceive themselves as greedy. Related to this, Fallis et al. (2015) report that the demand for goods which a share of the sales price is donated is increasing in this price. They also present evidence that this is due to the decrease in social image utility which consumer derive from purchasing such good-donation-bundles.

Prior research has also shown that when it comes to morally-ladden situations, people form self-serving assessments about what norms should apply and about others' preferences when it helps them to obtain a positive, moral self-image. Loewenstein et al. (1993) give subjects information about legal cases. These subjects then differ strongly in what they consider as appropriate, fair settlement values for these cases after they argued in fictitious roles of being the plaintiff as opposed to the defendant. Di Tella and Pérez-Truglia (2015) show evidence that people form beliefs about others behaving anti-socially, i.e. that others steal from a common pot, in order to justify their own anti-social behavior of not splitting the pot equally. People also employ uncertainty and ambiguity in a related manner to form self-serving beliefs and probability assessments which allow them to obfuscate their own immoral behavior (Dana et al., 2007; Haisley and Weber, 2010; Exley, 2016).

In this paper, we connect these findings to obtain insights about their lasting implications in the context of advice-giving. Closely related to our results is Gneezy et al. (2016): In several experiments, the authors show that advisers bias their recommendations relatively strongly when they learn about their conflict of interest before they receive the information about a client's decision situation. When they first learn about the situation, then consider what to recommend, and then about the conflict of interest, their advice is less biased. Following Trivers (2011), they label this behavior self-deception. Our theory and results describe behavior which is in line with such self-deception, i.e. that advisers effectively bias their own choices. We make the point that the reason for this behavior and the consistency in advisers' biased recommendation is that advisers try to avoid a negative self-inference.[3] This also relates to Konow (2000), who examines a dictator game where the pie to be split is dependent on the dictator's and the recipient's prior joint effort. He finds that dictators who allocate themselves larger shares of the pie interpret their personal contribution in establishing the common pot more favorably than outside observers.

[3]Falk and Zimmermann (2016) show that agents also act consistently to signal their skills to a principal. In Falk and Zimmermann (2015), they provide evidence that people act consistently without any external observers. The general idea which underlies the mechanism we propose also applies in these settings: Acting inconsistently shows that one's first action was somehow flawed, acting consistently therefore avoids such a inference to oneself and/or outside observers.

Documenting the persistence of such a self-serving bias, these dictators apply their persistent biased judgment about others' effort when they act as outside observers themselves.

Recent findings on actual advisers' behavior by Foerster et al. (2016) and Linnainmaa et al. (2016) relate to ours. Using matched data on about 5900 Canadian financial advisers and their more than 580,000 clients, these studies show that the most important determinants of advice to these clients are not the clients' personal characteristics, but rather the identity of their advisers. Even more important in our context, they show that these recommendations to clients are also reflected by the choices which advisers make for their own portfolios. For example, advisers prefer the return-chasing and actively managed funds they sell to clients also for themselves. This is puzzling since these investments do not perform better than the market. When fees are subtracted, clients' and their advisers' investments even significantly under-perform relative to the market. Our results and the theory we propose resonate with these findings. In addition, the experimental setup we use allows to abstract from concerns of advisers self-selecting into suitable environments which may drive such findings (e.g. risk-seeking advisers who choose to sell risky investments with sales commissions).

We identify a strong, causal, and lasting effect of bonuses in advice-giving. Our findings therefore contribute to the recent literature on the adverse effects of bonus payments (Agarwal and Itzhak, 2014; Bénabou and Tirole, 2016). We also point out the role of self-signaling in such a setting which connects directly to the recent research on the work culture and self-perception of those working in the financial industry (Cohn et al., 2014; Zingales, 2015). However, our findings apply also outside this specific financial context to advice on risky decisions more generally.

In the remainder of this paper, we present our findings in more detail. The next section describes a mechanism of how moral and self-image concerns can lead to persistent bias after advisers have faced a conflict of interest. Section 3 explains the design and procedures of the experiment in which we investigate this mechanism. Section 4 derives predictions and section 5 presents our results. Section 6 concludes by reviewing these results with respect to their implications for the economics of motivated beliefs, advice giving and its regulation. An appendix contains a formal model in which the predictions are derived; it also contains further data analysis and the experimental instructions.

2 Mechanism

In this section, we describe a behavioral mechanism in which advisers' concerns to appear impartial and moral can lead to the opposite behavior – a persistent bias in their advice. The

framework presented here also provides the assumptions that underlie a formal model which can be found in the appendix. To analyze an adviser's behavior, we assume his overall utility to depend on three parts: 1) consumption utility derived from monetary payoffs, 2) the moral cost of not giving impartial advice, and 3) diagnostic (dis-)utility of learning from actions which reveal that one's previous advice was biased.

While the first element of an adviser's overall utility is standard, the second reflects the fact that advisers might feel compelled, and often are, to act solely in a client's best interest. Not doing so then creates a moral cost. To determine when such a cost occurs, the question then arises what constitutes a "client's best interest", i.e. what constitutes impartial advice. We assume that an adviser can form a belief about his clients' preferences and therefore about the utility that clients experience when they follow his advice. Giving advice which does not maximize this assumed utility of the client would then be a violation of giving impartial advice and creates the moral cost. However, predicting others' preferences is inherently difficult. This applies in particular for risk preferences (Hsee and Weber, 1997; Eckel and Grossman, 2008; Harrison et al., 2013), even when the inference is conducted by trained financial advisers and there is no conflict of interest (Roth and Voskort, 2014). In the presence of external incentives which creates such a conflict, the uncertainty in estimating others' risk preferences can be instrumentalized in a self-serving manner: Advisers may form a belief about their clients' preferences such that their, potentially biased, advice is compatible with it.

However, there are limits to such self-serving beliefs. It is a robust psychological fact that people base their inferences about others' preferences on their own (Marks and Miller, 1987), in particular for risk preferences (Faro and Rottenstreich, 2006).[4] In consequence, advisers' own preferences also play a role in determining what is impartial advice. We capture this by assuming that advisers incur a moral cost when they recommend an option which they would not choose for themselves if they were in the client's position.

The third factor which matters for advisers is the diagnostic (dis-)utility they derive when they learn to have given biased advice, based on a model of self-signaling (Bodner and Prelec, 2003; Benabou and Tirole, 2011). In contrast to the moral cost of acting immorally, this dis-utility only occurs to an adviser *after* he has biased his advice, at the point when his later actions indicate exactly this fact to him. This can be captured by a dual-self model in which the "diagnostic self" of an adviser learns ex post about the other self's motive for giving advice,

[4]Though initially coined by Ross et al. (1977) as a "false consensus effect", the falsity of estimates of others' preferences based on one's own is not evident. Works by Hoch (1987) and Dawes (1990) demonstrate that such projection is not just statistically correct; they also show that people can often improve their accuracy in predicting others' preferences by relying more strongly on their own. Engelmann and Strobel (2000) show that subjects do so when they are incentivized to make accurate predictions.

e.g. whether prior advice was issued impartially and therefore was morally sound or whether it was corrupted. The important implication of such an inference is that advisers can only uphold a positive and self-serving belief of their prior motives for giving advice as long as they do not take actions which are incompatible with this.[5] Dual-self models have been used previously to explore how people infer about themselves, in particular their moral behavior (Benabou and Tirole, 2004; Grossman, 2015; Grossman and van der Weele, 2016). Here, we use it as a crucial device to describe the trade-off between keeping self-serving beliefs about one's own motives and taking contradictory actions.[6]

These three components together then have implications for how and, most importantly for how long, a conflict of interest affects advisers' choices and their recommendations. To see this, consider an adviser who issued a biased advice, thus an adviser whose pecuniary payoff for biasing advice outweighed his moral cost of doing so. If he is also concerned about the self-image, he then needs to continue to give the same biased advice again, especially when the conflict of interest has disappeared. The reason is that in order to later entertain the (counterfactual) idea that his initial advice was unbiased, it should be unaffected by the presence of an external incentive. Changing advice when the the conflict of interest disappears would then signal just the opposite. When an adviser's own preference stipulates what he should recommend to a client, this mechanism has even further consequences. This is because such a rule implies that in order to perceive oneself as unbiased, an adviser has to act according to his biased advice for himself.

In consequence, a behavioral trait which generally seems to be desirable, the preference to perceive oneself as a moral person, can lead to persistent biases in the context of advice giving. In addition, it can have a lasting effect on advisers' own choices to the degree that they assign diagnostic value to them. With this behavioral mechanism in mind, we set up the following experimental design to explore it in more detail.

3 Experimental design and procedures

At the beginning of the experiment, subjects were allocated to computer terminals in cubicles where instructions were shown to them on screen. Subjects acting as advisers were then informed that they would get GBP 5.00 as a show-up fee for participating in the experiment and that

[5] In essence, this reflects the desire to avoid cognitive dissonance (Festinger and Carlsmith, 1959) – a discrepancy between one's actions and one's beliefs about what is the norm one should follow (for economic models of cognitive dissonance, see Akerlof and Dickens (1982) and Rabin (1994)). For a discussion about how cognitive dissonance and motivated (self-)perception relate see also Kunda (1992).

[6] Apart from enabling us to capture this cognition, it also captures the fact that the inferring self "forgets" about the other self's motives. This is in line with research showing that people cannot perfectly recall their past decision motives nor foresee their future ones (Kahneman et al., 1997; Loewenstein and Schkade, 1999)

there would be further possibilities to earn money. They were also informed that they would act as advisers for clients who would be drawn from the same pool of subjects for a future experimental session and that clients would also receive the same show-up fee.

It was then explained to advisers that they would have to recommend which out of three investments, referred to as option A, B, and C, their clients should take. They were told that clients would only know that option A's payoff would depend more on luck than option C's while option B is intermediate in this regard. They were also told that clients would not know the options' payoffs or the associated probabilities. Advisers were informed that they, as advisers, would soon learn these exact parameters of the investment options before they had to make a recommendation.

The advisers' superior information was then given to them on a paper sheet which explained the three investment options in detail (for a copy of this sheet and the experimental interface see the appendix). The text on the sheet explained the following procedure of how an option's payoff was determined: After an option was chosen, a six-sided die would be rolled. Depending on the chosen option, this would then yield either a safe payoff or a lottery. This lottery was described as a (fair) coin toss with heads yielding GBP 20 and tails nothing. The following table which was also on that sheet summarizes how the die's result maps into these possible outcomes, depending on the chosen investment option:

Die equal to:	**Option A**	**Option B**	**Option C**
1 or 2	lottery: GBP 20 or 0	safe payment: GBP 12	safe payment: GBP 12
3 or 4	lottery: GBP 20 or 0	lottery: GBP 20 or 0	safe payment: GBP 8
5 or 6	lottery: GBP 20 or 0	lottery: GBP 20 or 0	lottery: GBP 20 or 0

Table 1: Description of the investment options as shown to advisers, "lottery" is a coin toss.

The text explained this procedure in detail and also contained several examples. Note that a choice among the three compound lotteries which these three options represent, allows to categorize the underlying risk preferences.[7] Comparing the differences between option A and B, only those who are willing to give up a safe payoff of GBP 12 to play a lottery with an expected payoff of GBP 10 instead, i.e. risk-seeking individuals, choose option A. Conversely, option C is preferred to option B only by those who want to sacrifice an expected payoff of GBP 10 for a safe payoff of GBP 8. Thus, only risk-averse individuals choose option C. Accordingly, Option

[7]This choice between possible sub-lotteries within a compound lottery is essentially a stripped-down version of a similar tasks used previously by Hsee and Weber (1997) and Holt and Laury (2002). For example, Holt and Laury (2002) let subjects choose ten times among ten pairs of lotteries and one of these choices is randomly chosen to be implemented. Over the ten pairs, each pair's second lottery is increasingly more risky than the first which allows to interpret the switching point between the first and second lottery as an indicator of risk preferences. We have essentially two such switching points (between A and B when the die equals 1 or 2 and between B and C when the die equals 3 or 4) which allows the categorization along risk-seeking/neutral/averse preferences.

B is chosen by individuals who are neither sufficiently risk-averse nor sufficiently risk-seeking. Reflecting this ordering based on risk-preferences we will henceforth, with slight abuse of the precise meaning, refer to option A/B/C as the "risk-seeking/neutral/averse option".

Step 1 – First recommendation R1:

After having studied the instructions and choice situations, advisers were asked to make a recommendation to clients. For this, they had to write the sentence "I recommend you to choose option A/B/C", depending on what they wanted to advise, on a piece of paper which had their cubicle number on it. They were instructed to put this recommendation into an envelope, close it, and then click on a button on their screen. The envelope was then collected by an experimenter and put into a box. Before they made their recommendations, they were told that at the end of the experiment, one of the envelopes would be randomly drawn from the box to be presented to a client and that the corresponding cubicle number would be read aloud. An adviser thus knew that he would eventually know whether his recommendation was chosen to a be shown to a client before she would then have to make a choice.

Step 2 – Own choice O:

After all advisers had written down their recommendation R1 and all envelopes were collected, they were informed that they would now have to choose an investment option for themselves. Advisers were previously not informed about this step. The procedure was the same as for issuing advice: Subjects had to write on a sheet "I choose option A/B/C." and then put it in an envelope. An experimenter came by and collected the envelopes and put it in a separate box. Again, they were informed that at the end of the experiment, one of the envelopes would be chosen randomly, its number would be announced aloud, and that the respective adviser would be asked later to roll the die to determine his chosen option's payoff. Ex-ante, the choice situation and its implementation probability was thus the same as the one on which they had previously advised a client on.

Step 3 – Second recommendation R2

After advisers made their own choice O, they were asked to make a second recommendation. The procedure was exactly the same as for R1, including the collection of envelopes in a separate box, sampling one from it and announcing its number. Again, advisers did not know in advance about this step. Advisers were also informed that this second advice, if it was sampled, would be shown to a different client in the same future session with clients.

Step 4 - Questionnaire and implementation:

After all recommendations were collected, subjects filled out a short questionnaire which elicited personal characteristics. The experimenter then sampled one envelope from each of the boxes

which contained the envelopes for R1, O, and R2 and announced the respective cubicle numbers. Subjects were then paid out in private based on whether they were offered a bonus and their recommendations; the subject in each session whose own choice O was sampled also rolled the die and received the corresponding payoff.

NO BONUS versus BONUS treatment: The above describes the experimental procedure in our baseline condition to which we will refer as NO BONUS. Our experimental manipulation was to offer some advisers a bonus for recommending the risk-seeking option A in R1. We will refer to this treatment as BONUS. After having been informed about the advice they had to give and how to do so, but before seeing the sheet with the detailed information about the investment options, every second adviser (in total 48) in a given session was randomly determined to be in that treatment. These advisers were informed that they would get a bonus of GBP 3 if they recommended option A. This bonus was only paid for subject's first recommendation R1. For those advisers who were offered the bonus, there were explicit notifications on the screens which explained the O and R2 tasks which clearly stated that there would not be any bonus for these tasks.[8] This within-session, across-subjects intervention with regard to the bonus is the only difference between our NO BONUS and BONUS.

Verifiability: In order to ensure that advisers believed that a recommendation, if randomly chosen to be shown to a client, would be actually seen by the client we allowed advisers to sign their recommendations and to address the envelopes to themselves. Advisers were explained that if their recommendation was chosen to be shown a client, the sheet would be signed by the respective client. In case that the corresponding adviser had provided us with his or her address, this subject would then get a copy of the signed recommendation by post. In addition, they were informed that this mailing would also contain information on how they could see the original, signed receipts which were deposited with the lab's official record depository. Subjects were informed of this before making their first recommendation. Since an adviser knew that he would know whether his envelope was sampled, this procedure pre-committed us to actually show the sampled advice letters to actual clients.

General procedures: Throughout the experiment, we enforced a strict no communication policy. We conducted eight sessions, each with 11 to 14, in total 99, subjects acting as advisers. Advisers earned on average GBP 6.68 ($9.51 at the time of the experiment) while no session lasted longer than 45 minutes. All subjects were students across several degrees and fields of

[8]Since advisers' payoff in BONUS do not depend on the clients' decisions, they were not explicitly informed about whether clients would learn about the bonus. Also none of the advisers asked for this information. In the session with clients, they were informed of the bonus when they received a recommendation R1 from an adviser who had been in the BONUS treatment.

studies. Table 12 in the appendix shows descriptive statistics. The experimental sessions were conducted in late January 2016 at the London School of Economics's Behavioural Research Lab with subjects from its pool. The experimental interface was implemented using zTree (Fischbacher, 2007). A week after the eight adviser sessions, we invited 16 additional subjects from the same pool for an additional session. In this session, they acted as clients and received the sampled recommendations from the previous adviser sessions, made their choices, and were paid their resulting payoffs. In this paper, we only focus on advisers and their recommendations.[9]

4 Predictions

In this section, we derive predictions for our experiment. They are based on the assumptions which we described in section 2, thus on advisers maximizing their overall utility from pecuniary payoffs, the moral cost of giving in-appropriate advice, and the self-image concern. Given our treatment intervention, we make the predictions with regards to how often the risk-seeking option A is recommended and chosen. All predictions derived and presented in this section are also derived in the formal, mathematical model which can be found in the appendix.

Predictions for R1: In NO BONUS, there is no pecuniary gain of issuing any specific recommendation. Since this is the first choice which an adviser makes it does not have signaling value with regards to past behavior. Absent pecuniary motives, only the moral cost of issuing inappropriate advice therefore remains. Beliefs about client's preference can be formed in a self-serving way, i.e. such that they suit an adviser's recommendation, up to the point that they contradict his own preference. To minimize the cost from recommending something that one would not choose for oneself, advisers thus recommend option A only if they prefer it. Thus, only risk-seeking adviser recommend option A.

In the BONUS treatment this is different: Advisers are now paid for recommending option A and derive pecuniary utility from the bonus when they do so. Clearly, those who would have recommended it anyhow, i.e. risk-seeking advisers, also recommend it in this treatment and in addition, get the bonus. However, those who would not have recommended it in the NO BONUS because they do not prefer it themselves now face a trade-off: When the moral cost of recommending something they would not choose for themselves are smaller than the pecuniary value of the bonus, they recommend option A. Otherwise, they recommend their

[9]With only 16 client observations which are not balanced over treatments (only three are eventually with recommendations from BONUS; recall that the probability of a recommendation being chosen is independent of the treatment), any analysis of client would have limited statistical power. However in the two experiments of Gneezy et al. (2016) which are in a related setting but have much more client observations, clients followed advisers in 74% and 85% of all cases, respectively.

preferred option. In both cases, they hold self-serving beliefs about the client's preference which is compatible with their issued advice. Assuming that some advisers have sufficiently low moral cost and follow the offered bonus, we get the following prediction:

Prediction 1. *There are more advisers in BONUS than in NO BONUS who recommend option A for the first recommendation R1.*

Predictions for O: In contrast to the first recommendation, advisers now make choices for themselves. The moral cost of giving inappropriate advice are therefore absent. Since the NO BONUS did not feature a bonus, there was no incentive to act immorally and to give biased advice. In consequence, there is no concern about drawing any (negative) inference from the own choice about one's preceding advice. The only relevant decision criterion is thus one's own risk preference and only risk-seeking advisers should choose option A for themselves in NO BONUS.

The own choice situation in BONUS and the NO BONUS is identical. Differences in behavior must occur because advisers in BONUS have previously been exposed to the bonus and, potentially, have given in to it. To the degree that they assign diagnostic value to their choices, advisers' own choices can then reveal to themselves that they were corrupted by the bonus: Advisers who recommended option A in R1 should, in order to appear as having given appropriate advice, also prefer it for themselves. In order to uphold the self-image that they were not corruptible, advisers who recommended option A just for the bonus must then mimic the incorruptible ones by choosing option A for themselves.[10] However, these advisers lose expected pecuniary utility because they choose the option which they do not actually prefer. In consequence, only those corruptible advisers who have sufficiently high image concerns, relative to their loss in expected pecuniary utility, choose option A for themselves, in addition to the incorruptible, risk-seeking ones. Note, however, that this only applies if own choices have sufficient diagnostic value, i.e. if advisers acknowledge the reverse implication of "I should recommend to my client what I would choose in her situation". Under the assumption that advisers assign such diagnostic value to their own choices we predict the following:

Prediction 2. *There are more advisers in BONUS than in NO BONUS who recommend option A for the the own choice O.*

[10] In terms of a signaling model, this is an equilibrium where corruptible advisers pool with those who truly prefer option A. In principle, there could be other equilibria where corruptible advisers and those who truly prefer option A pool on choosing non-A options, together with incorruptible advisers who actually prefer these options. However, in terms of self-signaling, these are rather unrealistic equilibria. This is so because in such equilibria, those who behaved morally obfuscate their behavior while those who behaved immorally do not. We therefore exclude them. We discuss this in more detail in the formal model in the appendix. There, we also show that these excluded equilibria do not even need to exist. In contrast, the former one where corruptible advisers mimic incorruptible ones by choosing option A does always exist.

Second recommendation R2: The predictions for the second recommendation combine insights from above. In NO BONUS, an adviser's pecuniary utility is unaffected by his second recommendation. Also, absent any previous bonus to give inappropriate advice, self-signaling concern do not play any role either. Accordingly, only the moral cost for giving inappropriate advice matters, as in R1. A previously formed self-serving belief coincides with the previous recommendation. For this recommendation, an adviser's own preference was the determining factor so that again, only risk-seeking advisers recommend option A (again).

In the BONUS treatment, the second recommendation does not entail any bonus either. However, the bonus which was offered to advisers in R1 opens the possibility that this recommendation was biased and therefore, the concern for signaling one's own corruptibility matters. Advisers who truly prefer option A can then minimize the moral cost of giving inappropriate advice and the self-signaling concern by recommending option A again in R2. As outlined above, advisers who do not prefer option A but recommended it in R1 for the bonus may mimic the incorruptible ones by choosing option A in O to prevent dis-utility from learning that they gave biased advice. Following the same logic, they can then mimic the incorruptible ones by re-recommending option A in R2. Note that the situations in R2 and R1 are identical, except for the bonus. Therefore, an inconsistency is more directly attributable to one's corruptibility; the second recommendation should have higher diagnostic value than the own choice. We thus get the following prediction:

Prediction 3. *There are more advisers in BONUS than in NO BONUS who recommend option A for the second recommendation R2.*

Conditional on a scenario in which at least some advisers are corrupted by the bonus, thus that prediction 1 is true, our design enables us to investigate two main questions. First, by testing prediction 3 we can find evidence for self-image concerns which cause repeated bias in advice-giving. If advisers are only steered by pecuniary incentives and not by the diagnostic value of their actions, we would not expect differences between BONUS and NO BONUS. In addition, comparing the own choice O across treatments allows to test whether they also have diagnostic value. If they do not, advisers should just implement their preferred choices which, due to random treatment assignment, should not differ between BONUS and NO BONUS. However, if prediction 2 is also confirmed, this indicates that advisers make choices which are, from a purely pecuniary point of view, sub-optimal just to appear incorruptible. It would therefore indicate that they assign diagnostic value to their own actions.

With this in mind, we will next examine the actual advisers' behavior in our experiment.

Before doing so, it is noteworthy that the proposed mechanism is, in principle, also capable of explaining the findings by Gneezy et al. (2016). They report on an experiment in which they expose advisers to a bonus and to a decision situation similar to ours. They then examine the effect of when this exposure to the bonus happens. They find that recommendations are less affected by the bonus when advisers learn about it after they have first considered what to recommend. In contrast, when they know about the bonus before such a consideration, their following advice is more biased. If the act of actively considering what to recommend also has diagnostic value, then changing one's actual recommendation afterward, once one has learned about the bonus, would also signal one's corruptibility. If in contrast an adviser knows from the beginning about the bonus, this can already be taken into account when initially considering what to recommend. He can then form a self-serving belief which supports his biased consideration and therefore also the actual recommendation. This would prevent a negative self-inference.

5 Results

Results for R1: This is where our treatment manipulation occurred. In the BONUS treatment, advisers were paid a bonus to recommend option A. Accordingly, we expect some to give in to this incentive and recommend it. This is also what we observe: In the NO BONUS only 3.9% of advisers recommend option A in their first recommendation. In contrast, about half of all advisers (54.2%) in BONUS recommend this option – an increase by 50.3 percentage points which is highly significant (Fisher exact test: $p = 0.000$).[11] Figure 1 shows the overall distribution of choices across these treatments:

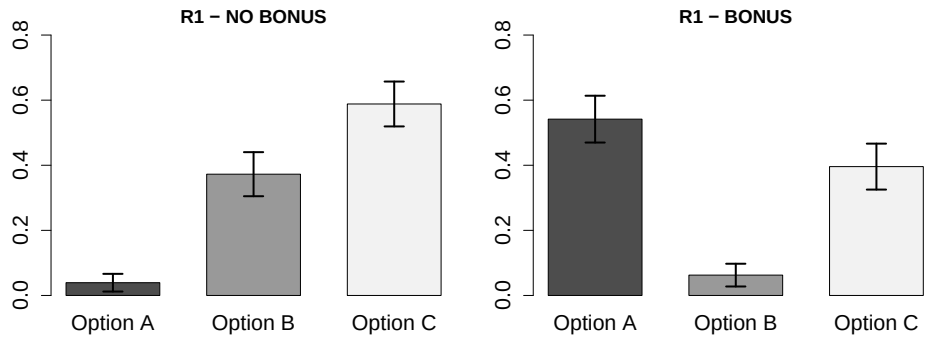


Figure 1: Frequency for each option being recommended in R1, bars depict standard errors.

We also employed a parametric approach via the following linear probability model which allows

[11] Although we have directed hypothesis, the reported p-value here and in the the following always refer to more conservative two-sided hypotheses.

us to control for the effect of remaining heterogeneity across treatments or sessions:

$$\text{Prob}[r_{1,i} = A] = \alpha + \beta \cdot BONUS_i + \boldsymbol{\delta} \cdot \mathbf{c}_i + \boldsymbol{\gamma} \cdot \mathbf{s}_i + \epsilon_i \tag{1}$$

In the above, $r_{1,i}$ is subject i's first recommendation out of the set of possible recommendations $\{A, B, C\}$ and $BONUS_i$ is a dummy indicating whether this subject was randomly assigned to the treatment BONUS. The vector $\mathbf{c}_i$ collects control variables which indicate a subject's age, gender, monthly budget, dummies for regions of origin, the highest degree a subject holds or pursues and his or her fields of studies. Control dummies for each session are collected in $\mathbf{s}_i$. The error term ϵ_i captures idiosyncratic noise in the decision for an adviser's recommendation. Table 2 presents the results when controls are successively added. It shows that the increase of about

	(1)	(2)	(3)	(4)
BONUS	0.502***	0.497***	0.489***	0.481***
	(0.078)	(0.076)	(0.093)	(0.092)
Personal Controls	no	yes	no	yes
Session Controls	no	no	yes	yes
Observations	99	99	99	99
Adjusted R^2	0.304	0.323	0.280	0.310

Table 2: OLS estimates of the probability to recommend option A in R1
robust standard error in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
personal controls: age, gender, monthly budget, subject's region of origin and field of studies

50 percentage points in the probability of recommending option A is almost unaffected by the addition of these controls and remains highly significant. We also repeat the same estimation procedure by probit and do not find any qualitative differences (see table 8 in the appendix). We therefore note that our treatment manipulation worked and that prediction 1 is confirmed.

It is also noteworthy that our results indicate that when offered a bonus, almost half of our subjects do *not* recommend option A. If subjects were confused or indecisive we would expect them all to take the money. However, there is something which stops a significant share, 45.8% (t-test: $p = 0.000$), of all advisers in BONUS from recommending this option, even for money. The notion of advisers refusing to recommend it because they consider it inappropriate or immoral advice is consistent with this observation.

Own choice O: For their own choice, no bonus is paid to advisers in both conditions. Figure 2 displays their choices. In the baseline NO BONUS we observe that 9.8% choose option A for themselves.[12] In BONUS however, when advisers were *previously* offered the bonus for their first

[12]This share is in the same region as the six to eight percent of subjects reported by Holt and Laury (2002) who exhibited risk-seeking behavior in a similar lottery without preceding advise.

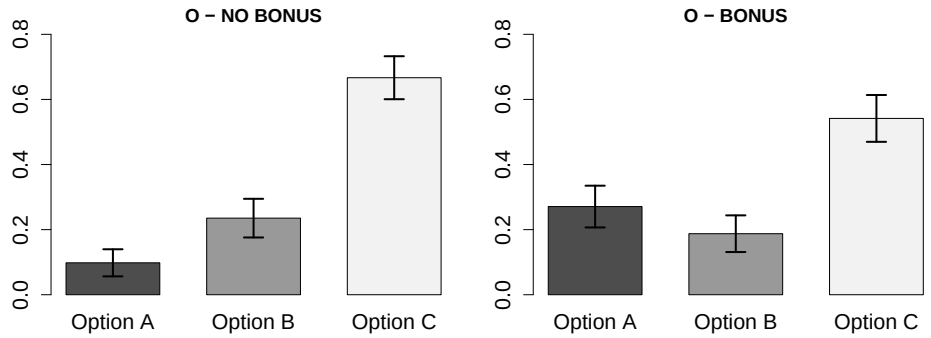


Figure 2: Frequency for each option being chosen in O, bars depict standard errors.

recommendation, 27.1% of all advisers, almost three times as much as in NO BONUS, choose the risk-seeking option A for themselves. This increase by 17.3 percentage point is significant (Fisher exact test: $p = 0.036$).

This finding is also confirmed when we re-estimate model (1) with a dummy indicating that an adviser chooses option O for himself as the dependent variable. Table 3 reports the corresponding results when the same control variables as in the preceding analysis are successively added. The effect of being in BONUS even increases and this pattern is again similar when the model is estimated by probit (see table 9 in the appendix). Therefore, we regard prediction 2 as confirmed.

	(1)	(2)	(3)	(4)
BONUS	0.173**	0.178**	0.219**	0.218**
	(0.077)	(0.081)	(0.095)	(0.087)
Personal Controls	no	yes	no	yes
Session Controls	no	no	yes	yes
Observations	99	99	99	99
Adjusted R^2	0.040	0.010	0.065	0.088

Table 3: OLS estimates of the probability to choose option A for oneself in O
robust standard error in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
personal controls: age, gender, monthly budget, subject's region of origin and field of studies

Given these findings, it is helpful to recall the mechanism which underlies our prediction since we can examine this causal channel more closely. The mechanism argues that if advisers assign diagnostic value to their own choice, they have to act according to their (biased) advice in order not to self-signal that they were corrupted. Our findings for R1 indicate that the bonus corrupted about half of all advisers; it leads to an increase of recommending option A by 50.3 percentage points for BONUS relative to NO BONUS. The findings on advisers' own choice O just presented, show that there is an increase of 17.3 percentage points for those who were potentially corrupted, i.e. those who were exposed to the bonus. These estimated

probabilities then imply the share of advisers who choose option A for themselves because they have previously given in to the bonus but do not want to self-signal their corruptibility is given by 34.4% ($\triangleq 0.173/0.503$).[13] This estimate shows that more than a third of those advisers who were put on the spot by biasing their recommendations and then having to choose for themselves behaved consistently by choosing option A for themselves.

We can also take our choice rate for option A in NO BONUS, which is 9.8%, as an estimate of how many people actually prefer it independent of possible image concerns due to the bonus. Adding this to the above estimate, we would expect that a total of 44.2%(=34.4%+9.8%) of the advisers in BONUS who initially recommended option A in R1 behaved consistently and also chose it in O. What we empirically observe is that 42.3% of the advisers in BONUS who initially recommended option A exhibit such a behavior, a percentage which is not different from the expected one (t-test: $p = 0.850$). Furthermore, this observed frequency also means that a significant share of advisers in treatment BONUS who have initially recommended option A, 57.7% (t-test: $p = 0.000$), do *not* choose it for themselves. Again, if advisers were just confused and took the bonus as an indication of what they should recommend, we would expect them all to also act accordingly for themselves.

Second recommendation R2: For their second recommendation, the decision situation for advisers in NO BONUS is the same as for their first. Accordingly, we expect a similar pattern of recommendations. The left panel of figure 3 shows the recommendation frequencies for each option. Comparing it to the left panel of figure 1 shows that this is largely the case: 82.4% of the advisers in NO BONUS recommend again exactly the same option they recommend initially. In particular, exactly the small minority of 3.9% of the advisers who recommended option A recommends it again.

This picture is very different when we compare this to the recommendations in BONUS. Although there is no bonus for recommending option A in R2 either, the rate of recommendation

[13]This follows from re-arranging the following: The observed increase between NO BONUS and BONUS in own choices for A (17.3%) has, according to the described mechanism, to equal advisers' propensity of feeling compelled to choose option A for themselves due to their previous recommendation for it, multiplied with the increase in the probability of them recommending option A as caused by the bonus (50.3%). To capture this effect in our regression framework, we implemented the following two-stage procedure: In the first stage, we took our regression results for (1) to obtain an estimate of how strongly the bonus lead advisers to recommend option A. To see how this causal channel affected their own choice, denoted by c_i, we then estimated in a second step the model $\text{Prob}[c_i = A] = \alpha + \beta \cdot \widehat{r_{1,i} = A} + \delta \cdot \mathbf{c}_i + \gamma \cdot \mathbf{s}_i + \epsilon_i$ where $\widehat{r_{1,i} = A}$ is the predicted probability of adviser i recommending option A because i is exposed to the bonus, thus we take $Bonus_i$ and our first-stage results to instrument $r_{1,i}$. The estimate for β in the second stage then reflects the causal effect of the bonus on the probability of choosing option A for oneself. The point estimates range from 0.344 to 0.452, depending on the specification, and are significant ($p < 0.05$). Strictly speaking, the results of this two-stage procedure may however be biased since the exclusion restriction for the instrument $Bonus_i$ could be violated (being in the BONUS treatment could influence the own choice via channels other than the first recommendation). Given the fit to our above estimates and observations, we however consider the results of this procedure noteworthy.

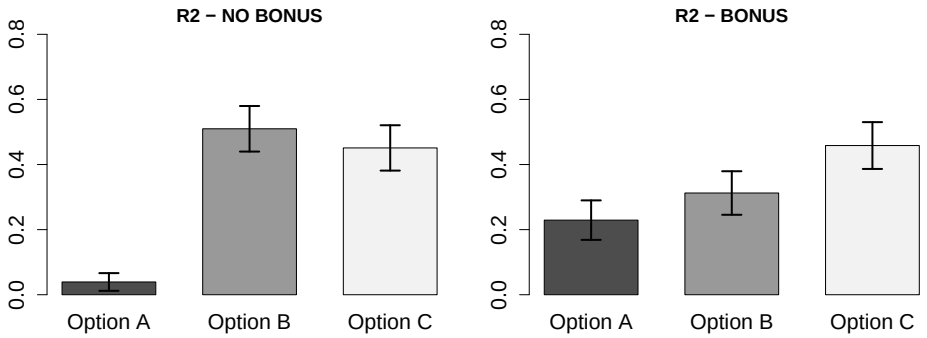


Figure 3: Frequency for each option being recommended in R1, bars depict standard errors.

for option A is almost five times as high as in the NO BONUS: 22.9% of those advisers who had previously been exposed to the bonus recommend option A, a significant increase by 19.0 percentage points relative to the NO BONUS (Fisher exact test: $p = 0.007$). This is also confirmed by a regression analysis which re-estimates model (1) when a dummy which indicates whether option A is recommended in the second recommendation is the dependent variable. Table 4 presents the results and shows that this point estimate even increases. Again, this

	(1)	(2)	(3)	(4)
BONUS	0.190***	0.203***	0.211**	0.213**
	(0.067)	(0.067)	(0.092)	(0.087)
Personal Controls	no	yes	no	yes
Session Controls	no	no	yes	yes
Observations	99	99	99	99
Adjusted R^2	0.070	0.073	0.038	0.064

Table 4: OLS estimates of the probability to recommend option A in R2
robust standard error in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
personal controls: age, gender, monthly budget, subject's region of origin and field of studies

pattern is also observed for probit estimates (see table 10 in the appendix). We therefore treat prediction 3 as confirmed.

As above for advisers' own choices O, we can estimate the causal effect of having given in to the bonus on the repeated recommendation for the risk-seeking option. The initial effect of an increase in the probability of recommending A in R1 due to the bonus was estimated by 50.3 percentage points. The observed increase of 19.0 percentage point in R2 then implies that, in expectation, 37.8% ($\triangleq$ 0.190/0.503) of advisers recommend option A again just because they have previously given in to the bonus.[14] To estimate the frequency of advisers in BONUS who

[14]We also repeated the two-step instrumental-variable-procedure as explained in footnote 13. That is, we estimate the probability of recommending option A again in R2 when one's first recommendation R1 has been biased the bonus. With the same caveat as described there applying here, the resulting IV-estimates of this causal channel range from 0.372 to 0.442 percentage point, depending on the specification, and are significant ($p < 0.01$).

recommend option A twice we add the 3.9% who do so in the NO BONUS treatment as an estimate for the proportion of those who recommend it for reasons unrelated to the bonus. The implied point estimate from this decomposition is 41.7%(=37.8%+3.9%). This estimate in the region of the actually observed frequency of advisers in BONUS who re-recommend option A in R2: It is given by 34.6% which is not statistically differed from the above estimate (t-test: $p = 0.417$).

Further results: There are some further findings which support our theory and its underlying assumptions. Given our previous results, we expect high consistency between advisers' own choices and their first recommendation when there is no conflict of interest. Our results are largely in line with this: Table 5a) shows the frequencies of advisers choosing for themselves, conditional on their first recommendation in NO BONUS. Only the off-diagonal entries are not in line with this prediction. They amount to a total of 17.7% of the observation in this treatment; 82.3% of our observations in NO BONUS are therefore in line with the predicted consistency. In BONUS, our theory predicts that some of those who have previously recommended option

		O = A	O = B	O = C
R1 =	A	3.9%	0.0%	0.0%
	B	2.0%	23.5%	11.8%
	C	3.9%	0.0%	54.9

a) NO BONUS

		O = A	O = B	O = C
R1 =	A	22.9%	8.3%	22.9%
	B	0.0%	6.3%	0.0%
	C	4.2%	4.2%	31.3

b) BONUS

Table 5: Frequencies of advisers' own choices O conditional on their first recommendation R1.

A stick to it in order to avoid a negative self-image. Other advisers who have recommended it but who do not have sufficiently strong image concerns choose their preferred option instead. Accordingly, we can explain the diagonal entries in table 5b) plus the off-diagonal ones in the first row. Again, this leaves only a small fraction of 8.4% of our observations unexplained.

We find similar results with regards to the consistency between advisers' first and second recommendations. Table 6a) and b) show the respective conditional frequencies across our experimental conditions. In NO BONUS, noise is somewhat higher than for the previous comparison. We observe a total of 21.6% to be inconsistent, i.e. to be outside table 6a)'s diagonal. However, one should note firstly, that these inconsistencies are primarily due to switches from having initially recommended option C and then option B, thus between neighboring, non-risk-seeking options. Secondly, almost eighty percent of recommendations are consistent and thus in line with our theory. With regard to variations in the BONUS treatment the results are even stronger. In total, 87.5% of our observations fall into an explainable pattern, thus are either on

		R2 =		
		A	B	C
R1 =	A	3.9%	0.0%	0.0%
	B	0.0%	35.3%	2.0%
	C	3.9%	15.7%	43.1%

a) NO BONUS

		R2 =		
		A	B	C
R1 =	A	18.8%	16.7%	18.8%
	B	0.0%	6.3%	0.0%
	C	4.2%	8.3%	27.1%

b) BONUS

Table 6: Frequencies of advisers' second recommendations R2 conditional on their first R1.

the diagonal or the first row. Overall, the consistency predicted by our theory can be observed in at least four fifth of the relevant cases and often, in even higher proportions.

Further evidence comes from our exit questionnaire. It contained a question on advisers' general risk attitudes. More precisely, it asked subjects to indicate on an 11-point Likert-scale "How willing are you to take risk, in general?". Although this question was not incentivized, answers to it has previously been shown to correlate with peoples' actual choices under risk (Dohmen et al., 2011). While in NO BONUS, the average response was 5.0 points, it increased by almost one point or alternatively, 39.8% of its standard deviation, to 5.9 points in the treatment BONUS. This increase is marginally statistically significant (Wilcoxon ranksum-test: $p = 0.059$).[15] This result becomes even stronger, both numerically and statistically, in an OLS regression analysis when additional control variables are included. Table 7 represents the results from estimating model (1) when the dependent variable is this self-assessed risk-measure and controls are successively added. The results are also robust to estimation via ordered probit (see table 11 in the appendix). This increase in an adviser's self-stated risk measure is consistent with our theory: Advisers who have previously given in to the bonus can self signal that this advice was appropriate from their point of view when they consider themselves as more risk-seeking.[16] Once again, this is also consistent with advisers who are not just confused about their choices and recommendations but who, on the contrary, do even understand the more general behavioral implications of their recommendations outside the given set of options.

[15] Due to a data-glitch in the first two sessions, we had to collect the risk-measure along with the other post-experimental questionnaire separately. When we exclude these sessions, the increase is 1.1 points, 46% of the measure's standard deviation, and is similarly significant (Wilcoxon ranksum-test: $p = 0.062$.). The same pattern (higher point estimates and slightly lower but still significant p-values) holds when we exclude these observations from the regressions reported in table 7. Note that our primary data on the recommendations R1/R2 and own choices O were not affected by this data glitch since they were collected by advisers writing them on paper.

[16] We also repeated the two step instrumental variable procedure laid out along with its caveats in footnote 13. This allows us to estimate the effect on the risk measure through having recommended option A by instrumenting this choice via an advisers' random exposure to the bonus. The estimated coefficient ranges from a 1.8 to 2.2, depending on the specification and are significant ($p < 0.05$). Given the first stage increase in the probability of recommending option A due to the bonus of 50.3 percentage points, the implied causal increase of 0.9 to 1.1 (1.8×0.503 to 2.2×0.503) is consistent with these estimates.

	(1)	(2)	(3)	(4)
BONUS	0.914**	1.030**	1.244**	1.306**
	(0.453)	(0.436)	(0.576)	(0.590)
Constant	4.961***	3.534***	5.185**	5.819*
	(0.335)	(0.634)	(2.284)	(3.090)
Personal Controls	no	yes	no	yes
Session Controls	no	no	yes	yes
Observations	99	99	99	99
Adjusted R^2	0.040	0.199	0.374	0.415

Table 7: OLS estimates on the self-assessed preference for risk (Likert scale, 0 to 10)
robust standard error in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
personal controls: age, gender, monthly budget, subject's region of origin and field of studies

6 Conclusion

In this paper, we provide experimental evidence that incentives to bias advice have a lasting and causal effect on both, advisers' future recommendations for risky decisions and their own choices. When advisers are paid a bonus to recommend an investment option which is only preferred by risk-seeking individuals, about half of them recommend it. Without such a bonus only four percent do so. Prior exposure to a bonus leads a significant share of advisers to re-recommend this option when there is no bonus anymore and even to choose it for themselves. We provide a psychological mechanism which is capable of explaining these findings. It is based on advisers' desire to not self-signal their corruptibility. This forces them to be consistent in their recommendations and own choices, even when this means to bias further advice and even their own choices. With this theory we can consistently decompose the recommendation and choice pattern of advisers in our experiment. We estimate that around 35 to 40 percent of those advisers whose advice has been corrupted by the bonus engage in such continuing deception of advisers and also of themselves in order to preserve a positive self-image.

A straightforward policy implication of our findings is therefore that removing advisers' conflicts of interest does not necessarily eliminate their effect on advice giving. For example, the Retail Distribution Review (RDR) in the UK whose stepwise implementation started in 2013 bans commission-based financial advice. Our results indicate that, while it may improve such advice in the long run, its full effects may be considerably delayed. Experienced advisers, who have spent their hitherto professional life in an environment which featured such incentives, will likely exhibit persistent biases in their recommendations.

Our proposed mechanism also has profound consequences on how accountable advisers feel. It implies that it is the desire to see oneself as a moral, impartial adviser which can lead to exactly

the opposite behavior. Those who stop giving biased advice after bonuses are removed identify themselves as having previously been corrupted. In contrast, those who continue to give biased advice do so just to avoid this inference and therfore, do not feel corrupted. In consequence, the awareness of acting in a corrupted manner and actually giving biased advice do not coincide, in fact they are asymmetric. This provides challenges for the remedy of the biases resulting from conflicts of interest as those who do the damage might not even feel culpable. Given the demand for advice in many situations, we think that exploring these mental processes by advisers and the adverse consequences it has on their job is a fruitful avenue for further research.

References

Agarwal, S. and B. D. Itzhak (2014). Loan prospecting and the loss of soft information. *NBER Working Paper Series 19945*.

Akerlof, G. A. and W. T. Dickens (1982). The Economic Consequences of Cognitive Dissonance. *American Economic Review 71*(3), 437–447.

Benabou, R. and J. Tirole (2004). Willpower and Personal Rules. *Journal of Political Economy 112*(4), 848–886.

Benabou, R. and J. Tirole (2011). Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics 126*(2), 805–855.

Bénabou, R. and J. Tirole (2016). Bonus Culture: Competitive Pay, Screening, and Multitasking. *Journal of Political Economy*, forthcoming.

Bodner, R. and D. Prelec (2003). Self-Signaling and Diagnostic Utility in Everyday Decision Making. In I. Brocas and J. D. Carrillo (Eds.), *The Psychology of Economic Decisions Vol.1*, Volume 1, pp. 105–126. Oxford University Press.

Cain, D. M. and A. S. Detsky (2008). Everyone's a Little Bit Biased (Even Physicians). *Journal of the American Medical Association (JAMA) 299*(24), 2893–2895.

Cohn, A., E. Fehr, and M. A. Maréchal (2014). Business culture and dishonesty in the banking industry. *Nature* (516), 86–89.

Dana, J. and G. Loewenstein (2003). A social science perspective on gifts to physicians from industry. *Journal of the American Medical Association (JAMA) 290*(2), 252–255.

Dana, J., R. A. Weber, and J. X. Kuang (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory 33*(1), 67–80.

Dawes, R. (1990). The potential nonfalsity of the false consens effect. In R. M. Hogarth (Ed.), *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, pp. 179–199. University of Chicago Press.

Di Tella, R. D. and R. Pérez-Truglia (2015). Conveniently Upset: Avoiding Altruism by Distorting Beliefs About Others. *American Economic Review 105*(11), 3416–3442.

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*(3), 522–550.

Eckel, C. C. and P. J. Grossman (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior and Organization 68*(1), 1–17.

Engelmann, D. and M. Strobel (2000). The false consensus effect disappears if representative information and monetary incentives are given. *Experimental Economics 3*(3), 241–260.

Exley, C. L. (2016). Excusing Selfishness in Charitable Giving: The Role of Risk. *Review of Economic Studies*, forthcoming.

Falk, A. and F. Zimmermann (2015). Information Processing and Commitment. *mimeo*.

Falk, A. and F. Zimmermann (2016). Consistency as a Signal of Skills. *Management Science*, forthcoming.

Fallis, A., X. Luo, and Z. Fang (2015). Self-Signaling and Prosocial Behavior: A Cause Marketing Mobile Field Experiment. *NBER Working Paper Series 21475*.

Faro, D. and Y. Rottenstreich (2006). Affect, Empathy, and Regressive Mispredictions of Others' Preferences Under Risk. *Management Science 52*(4), 529–541.

Festinger, L. and J. M. Carlsmith (1959). Cognitive consequences of forced compliance. *Journal of Abnormal Psychology 58*(2), 203–210.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Foerster, S., J. T. Linnainmaa, B. T. Melzer, and A. Previtero (2016). Retail Financial Advice: Does One Size Fit All? *Journal of Finance*, forthcoming.

Gneezy, A., U. Gneezy, G. Riener, and L. D. Nelson (2012). Pay-what-you-want, identity, and self-signaling in markets. *Proceedings of the National Academy of Sciences 109*(19), 7236–7240.

Gneezy, U., S. Saccardo, M. Serra-Garcia, and R. van Veldhuizen (2016). Motivated Self-deception and Unethical Behavior. *mimeo*.

Grossman, Z. (2015). Self-Signaling and Social-Signaling in Giving. *Journal of Economic Behavior & Organization 117*, 26–39.

Grossman, Z. and J. van der Weele (2016). Self-Image and Willful Ignorance in Social Decisions. *Journal of the European Economic Association*, forthcoming.

Haisley, E. C. and R. A. Weber (2010). Self-serving interpretations of ambiguity in other-regarding behavior. *Games and Economic Behavior 68*(2), 614–625.

Harrison, G. W., M. I. Lau, E. E. Rutström, and M. Tarazona-Gómez (2013). Preferences over social risk. *Oxford Economic Papers 65*(1), 25–46.

Hilgartner, S. (2000). *Science on stage: Expert advice as public drama*. Stanford University Press.

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology 53*(2), 221–234.

Holt, C. and S. Laury (2002). Risk aversion and incentive effects. *The American Economic Review 92*(5), 1644–1655.

Hsee, C. K. and E. U. Weber (1997). A fundamental prediction error: self-other discrepancies in risk preference. *Journal of Experimental Psychology: General 126*(1), 45–53.

Kahneman, D., P. P. Wakker, and R. Sarin (1997). Back to Bentham? Explorations of Experienced Utility. *The Quarterly Journal of Economics 112*(2), 375–405.

Konow, J. (2000). Fair Shares : Accountability and Cognitive Dissonance in Allocation Decisions Fair Shares. *American Economic Review 90*(4), 1072–1091.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin 108*(3), 480–498.

Kunda, Z. (1992). Can Dissonance Theory Do It All? *Psychological Inquiry 4*(3), 337–339.

Linnainmaa, J. T., B. T. Melzer, and A. Previtero (2016). The Misguided Beliefs of Financial Advisors. *mimeo*.

Loewenstein, G., S. Issacharoff, C. Camerer, L. Babcock, C. Camerer, and L. Babcock (1993). Self-Serving Assessments of Fairness and Pretrial Bargaining. *The Journal of Legal Studies 22*(1), 135–159.

Loewenstein, G. and D. Schkade (1999). Wouldn't It Be Nice? Predicting Future Feelings. In D. Kahneman, E. Diener, and N. Schwarz (Eds.), *Well-being: The foundations of hedonic psychology*, Chapter 5, pp. 85–105. Russell Sage Foundation.

Malmendier, U. and D. Shanthikumar (2014). Do security analysts speak in two tongues? *Review of Financial Studies 27*(5), 1287–1322.

Marks, G. and N. Miller (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin 102*(1), 72–90.

Mazar, N., O. Amir, and D. Ariely (2008). The Dishonesty of Honest People : A Theory of Self-Concept Maintenance. *Journal of Marketing Research XLV*(6), 633–644.

Mullainathan, S., M. Noeth, and A. Schoar (2012). The Market For Financial Advice. An Audit Study. *NBER Working Paper Series 17929*.

Mullen, B., J. L. Atkins, D. S. Champion, C. Edwards, D. Hardy, J. E. Story, and M. Vanderklok (1985). The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology 21*(3), 262–283.

Rabin, M. (1994). Cognitive dissonance and social change. *Journal of Economic Behavior & Organization 23*, 177–194.

Ross, L., D. Greene, and P. House (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology 13*(3), 279–301.

Roth, B. and A. Voskort (2014). Stereotypes and false consensus: How financial professionals predict risk preferences. *Journal of Economic Behavior and Organization 107*, 553–565.

SEC (2011). Study on Investment Advisers and Broker-Dealers. *Report by the staff of the U.S. Securities and Exchange Commission*.

Taylor, R. and J. Giles (2005). Cash interests taint drug advice. *Nature 437*, 1070–1071.

Trivers, R. (2011). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books.

United States Congress (2010). Dodd-Frank Wall Street Reform and Consumer Protection Act.

Zingales, L. (2015). Presidential Address: Does Finance Benefit Society? *Journal of Finance 70*(4), 1327–1363.

Appendix A – A simple model of self-signaling and corrupted advice-giving

In the following, we formally derive three predictions I to III which are analogous to their respective counterparts, predictions 1 to 3 in then main text. These derivations are based on a formal model presented below with assumptions capturing those described in section 2.

First recommendation R1: We consider an adviser who recommends a client which action out of a discrete set $\mathcal{S}$ to take. In our experiment, these are three investment options A, B, and C, thus $\mathcal{S} = \{A, B, C\}$. We denote an adviser's (first) recommendation by $r_1 \in \mathcal{S}$. In addition, there is a bonus $b(r_1)$ which depends on the issued recommendation. In our experiment, an adviser gets a bonus b if he recommends option A, otherwise he does not get any bonus. We thus have $b(r_1) = b \cdot \mathbb{1}[r_1 = A]$. We denote the utility which advisers get from a pecuniary payoff x by the strictly increasing vNM-utility function $u(x)$.

In addition, an adviser i suffers dis-utility $k_i > 0$ to the extend that he recommends an option which is not in the client's best interest. What constitutes a client's best interest is based on two factors: First, it is the choice c^* which the adviser would make if he had to make the client's decision for himself, thus $c^* = \arg\max_{c \in \mathcal{S}} \mathrm{E}[u(c)]$. Second, we allow the adviser to hold a (motivated) belief about the client's preferences. This is captured by the vNM-utility function $\tilde{u}$ which denotes the adviser's belief about the client's preference. We can then denote the implied optimal choice, based on this first-order belief, by $\tilde{c}^* = \arg\max_{c \in \mathcal{S}} \mathrm{E}[\tilde{u}(c)]$. We let $\gamma \in [0, 1]$ denote the weight which advisers assign to their own preference in determining what it is the client's best interest as opposed to optimal recommendations based on their first-order beliefs about the client's preferences. An adviser's overall utility of recommending r_1 is then given by the following expression:

$$\tilde{V}(r_1) = u(b \cdot \mathbb{1}[r_1 = A]) - k_i\big(\, \gamma \cdot \mathbb{1}[r_1 \neq c^*] + (1 - \gamma) \cdot \mathbb{1}[r_1 \neq \tilde{c}^*]\big) \quad (2)$$

This allows several interpretations: When $\gamma = 1$, the question of what constitutes appropriate, morally sound advice is the same as "What would I choose if I were in the client's position?". Conversely, $\gamma = 0$ means that only what an adviser beliefs about others' preferences, not his personal consideration, is relevant for issuing appropriate advice. Values of γ within the unit interval can represent situations in between or when an adviser believes that a client has utility represented by u with probability γ and otherwise represented by $\tilde{u}$. The magnitude of k_i then scales concerns about issuing unsuited advice relative to pecuniary payoffs.

Advisers can form a belief about the client's preferences in a self-serving manner. That is, whenever they issue a recommendation r_1 they can maximize their overall utility by self-servingly believe that the clients' preferences $\tilde{u}$ are such that $\tilde{c}^* = r_1$. In this regard, γ can also be interpreted as how far such a self-serving belief can be formed, independently of and adviser's own preferences. Therefore, the recommendation r_1 which maximizes (2) is the maximizer of the following, more simple, expression:

$$v(r_1) = u(b \cdot \mathbb{1}[r_1 = a)]) - \gamma k_i \cdot \mathbb{1}[r_1 \neq c^*] \tag{3}$$

We let K_{c^*} denote the cdf of the distribution of an adviser's moral cost k_i, conditional on this adviser preferring option c^*, e.g. $K_A(x) = \Pr[k_i \leq x | c^* = A]$. For simplicity, we assume that each of these conditionals cdf's has pdf which is strictly positive over its support.[17] We also let $\alpha_{c^*} > 0$ denote the share in the population of advisers who have preferred action $c^* \in \mathcal{S}$.[18] For easier notification, we let $\alpha = \alpha_A$, i.e. in our experiment α is the share of advisers who are sufficiently risk-seeking to choose option A. We assume the above distributions and parameters to be common knowledge.[19]

R1 – NO BONUS: Since there is no incentive to bias advice, only the second part of (3) matters. This is maximized by $r_1^* = c^*$. In consequence, the share of advisers who recommend option A equals α.

R1 – BONUS: For those who have $c^* = A$, it follows from (3) that they should also recommend it. For those with $c^* \neq A$, they can either recommend option A nevertheless to earn the bonus or they recommend their preferred option $c^* \neq A$ and obtain a utility of $u(0)$. Advisers who do not prefer option A then recommend it if and only if $\gamma k_i < u(b) - u(0)$. By using the convention that $K_{c^*}\left(\frac{u(b)-u(0)}{\gamma}\right)\big|_{\gamma=0} = \lim_{x\to+\infty} K_{c^*}(x) = 1$ we can then define the following

$$\beta \equiv \sum_{c^* \in \mathcal{S}\setminus\{A\}} \alpha_{c^*} K_{c^*}\left(\frac{u(b)-u(0)}{\gamma}\right) = \alpha_B K_B\left(\frac{u(b)-u(0)}{\gamma}\right) + \alpha_c K_c\left(\frac{u(b)-u(0)}{\gamma}\right) > 0$$

Thus with a bonus, a share β of advisers is corrupted by the bonus and recommends option A, in addition to the share α who would have recommended this option anyhow.

Given the same expected population of advisers across BONUS and NO BONUS, as achieved

[17] Results do not change when the cdfs are allowed to be partially non-increasing, as long as at least one of the pdfs has some mass on sufficiently low values, i.e. that $K_{c^*}\left(\min\{u(b)-u(0), \mathrm{E}[u(c^*)-u(A)]\}\right) > 0$ for at least one $c^* \in \mathcal{S} \setminus \{A\}$.

[18] In consequence, the unconditional cdf $\Pr[k_i \leq x]$ is given by $\sum_{c^* \in \mathcal{S}} \alpha_{c^*} K_{c^*}(x)$.

[19] Note that when the signaling concern refers to a dual-self model where advisers ex-post infer their own type from actions, this common prior only refers to these selves. A common prior between individuals is not required.

by random treatment assignment, we can then state the following:

Prediction I. $\Pr[r_1 = A \mid \text{bonus }] = \alpha + \beta > \Pr[r_1 = A \mid \text{no bonus }] = \alpha$

It will be helpful to categorize advisers along three behavioral types $\theta \in \{1, 2, 3\}$. These types reflect the motives underlying their recommendation r_1 as follows:

Type 1 ($\theta = 1$): Advisers who have $c^* = A$ and recommend $r_1 = c^*$, share α.

Type 2 ($\theta = 2$): Advisers who have $c^* \neq A$ but recommend $r_1 \neq c^*$, share β.

Type 3 ($\theta = 3$): Advisers who have $c^* \neq A$ and recommend $r_1 = c^*$, share $1 - \alpha - \beta$.

Type 1 and 3 advisers give the same advice they would have given had the bonus been absent. Type-2-advisers are corrupted: They recommend option A not because they prefer it but because they were paid to do so. Note that this above categorization of types also applies in the NO BONUS-treatment, the respective shares however differ: Share α also recommends option A without a bonus. Type-2-advisers do not exist in this treatment thus we can treat β as if it were equal to zero and the share of type-3-advisers is given by $1 - \alpha$.

Own choice O: The extent to which advisers take their own choice as a "diagnosis" of the moral type in R1 is given by $\lambda \geq 0$. A value $\lambda \in (0, 1)$ would reflect that choosing for one-self is not exactly the same as recommending to others but also that is not unrelated; $\lambda = \gamma$ is then a natural case. In general, we assume that λ is some increasing function Λ of γ with $\lambda = \Lambda(\gamma) = 0$ if and only if $\gamma = 0$. This means that own choices only have diagnostic value to assess an adviser's previous recommendation when his own preference is, at least partly, relevant for issuing appropriate advice.

When λ is positive, an adviser's own choice $c \in S$ signals his underlying motives for his previous recommendation in R1. In particular, an adviser can potentially infer that he was a type-2-adviser according to the above classification. The cost of inferring that one is such a type, thus that one-seld is corruptible yield image dis-utility $l_i > 0$. By denoting the expected utility from choosing a lottery $c \in \mathcal{S}$ by $\mathrm{E}[u(c)]$, the overall utility of advisers is then given by

$$V(c|r_1) = \mathrm{E[u(c)]} - \lambda l_i \cdot \Pr[\theta = 2|r_1, c] \tag{4}$$

As before, we assume that l_i can be described by a family of commonly known conditional cdfs $(L_{c^*})_{c^* \in \mathcal{S}}$, e.g. $L_A(x) = \Pr[l_i \leq x | c^* = A]$.[20]

[20]This effectively constitutes a intrapersonal signaling game where an adviser of type (k_i, l_i) sends a message $(c|r_1)$ and then gets dis-utility when he infers from this that his type is such the he behaves according to the behavioral type $\theta = 2$.

O – NO BONUS: When there was no prior bonus, there are no type-2 advisers. In consequence, $\Pr[\theta = 2|r_1, c] \leq \Pr[\theta = 2] = 0$ holds and $c = c^*$ maximizes (4) via $\mathrm{E}[u(c)]$. The share of advisers choosing option A for themselves is thus given by α.

O – BONUS: We start with the case that $\lambda > 0$. First note that type-3-advisers who have previously recommended $r_1 \neq A$ cannot infer to be type-2-advisers, i.e. $\Pr[\theta = 2|r_1 \neq A, c] = 0$. All type-3-advisers therefore choose $c = r_1 = c^* \neq A$ to maximize (4). Type-1 and type-2 advisers can however both infer to be type-2 and would then suffer dis-utility l_i because they have the same initial recommendation $r_1 = A$. Denote the likelihood that a type-1-adviser chooses $c = A$ with $\tau_c = \Pr[c = A|\theta = 1]$ and that a type-2-adviser makes the same choice with $\pi_c = \Pr[c = A|\theta = 2]$. One then gets the following for the corresponding posteriors:

$$\Pr[\theta = 2|c = A, r_1 = A] = \frac{\pi_c \cdot \beta}{\tau_c \cdot \alpha + \pi_c \cdot \beta} \tag{5}$$

$$\Pr[\theta = 2|c \neq A, r_1 = A] = \frac{(1 - \pi_c) \cdot \beta}{(1 - \tau_c) \cdot \alpha + (1 - \pi_c) \cdot \beta} \tag{6}$$

It is easily verified that $\Pr[\theta = 2|c \neq A, r_1 = A] \geq \Pr[\theta = 2|c = A, r_1 = A]$ whenever $\tau_c \geq \pi_c$. If this condition holds, type-1-advisers who choose $c \neq A$ suffer for two reasons: First, they loose expected pecuniary utility by choosing a suboptimal choice $c = A \neq c^*$. Second, they expect dis-utility from damage to self-image which is at least as big as when they had chosen their preferred option. In consequence, there is only one equilibrium with $\tau_c \geq \pi_c$ in which $\tau_c = 1$ and all type-1-advisers are consistent by choosing $r_1 = c = A$. While other equilibria with $\tau_c < \pi_c$ cannot be excluded but also do not need to exist, the one with $\tau_c = 1$ is a natural candidate: In it, type-1-advisers who are not corrupted by the bonus do also not deviate from their preferred choice just because of the fear of perceiving themselves as corruptible type-2-advisers while type-2-adviser, who want to uphold a positive self-image, might do so. Also, while there is always the equilibrium with $\tau_c = 1$, those with $\tau_c < \pi_c$ may not even exist.[21]

Type 2-advisers then face a trade-off: They would not like to choose option A for themselves, since for them $c^* \neq A$holds. However, if they switch from their first recommendation to their preferred option, they then generate a perfect signal of being type-2 since all other types are consistent by choosing $c = r_1$ and therefore, $\Pr[\theta = 2|c \neq A, r_1 = A] = 1$ holds. Using the posterior (5), a type-2-adviser therefore chooses his preferred option $c^* \neq A$ if and only if

$$\mathrm{E}[u(c^*)] - \lambda l_i > \mathrm{E}[u(A)] - \lambda l_i \cdot \frac{\pi_c \cdot \beta}{\alpha + \pi_c \cdot \beta}$$

[21] If the dis-utility of not choosing option A although one prefers it is too large, type-1 would not choose another option just to appear less as a type 2.

That is, an adviser reveals himself when his image concern is sufficiently low, i.e. when $l_i < \frac{\alpha+\pi_c\beta}{\lambda\alpha}\left(\mathrm{E}[u(c^*)] - \mathrm{E}[u(A)]\right)$. For this, they have to take into account that by not choosing option A, they decrease π_c. This in turn simplifies pooling and thereby raises the opportunity cost of such a choice. It follows that, in equilibrium, the share of type-2-advisers who choose option A to uphold a positive image balance this effect. This share is therefore given by the solution to the following expression:

$$\begin{aligned} 1-\pi_c &= \sum_{c^*\in\mathcal{S}\setminus\{A\}} \alpha_{c^*} L_{c^*}\left(\frac{\alpha+\pi_c\beta}{\lambda\alpha}\left(\mathrm{E}[u(c^*)] - \mathrm{E}[u(A)]\right)\right) \\ &= \alpha_B L_B\left(\frac{\alpha+\pi_c\beta}{\lambda\alpha}\left(\mathrm{E}[u(B)] - \mathrm{E}[u(A)]\right)\right) + \alpha_C L_C\left(\frac{\alpha+\pi_c\beta}{\lambda\alpha}\left(\mathrm{E}[u(C)] - \mathrm{E}[u(A)]\right)\right) \end{aligned} \tag{7}$$

Note that for all values of $\pi_c \in [0,1]$, the above RHS is strictly positive and non-decreasing in π_c. Also note that from $\alpha = \alpha_A > 0$ it holds that $\sum_{c^*\in\mathcal{S}\setminus\{A\}} \alpha_{c^*} L_{c^*}(x) < \sum_{c^*\in\mathcal{S}} \alpha_{c^*} L_{c^*}(x) \leq 1$ for every $x \in \mathbb{R}_{++}$.[22] The above RHS is therefore strictly less than one. Since the LHS is strictly decreasing in π_C and takes all values in $[0,1]$ over that interval, there is a unique solution $\pi_c^* \in (0,1)$ to (12). Also note that since the RHS of (12) is decreasing in λ, the implied consistency in own choice π_c^* is also strictly increasing in this parameter.

Now consider $\lambda = 0$: The second part in (4) does not count then and irrespective of their prior behavior, all advisers choose c^*. This is equivalent to $\pi_c^* = 0$.

In summary, share α of type-1-advisers initially recommend and then choose for themselves option A. Type-3-advisers initially recommend and then choose their preferred non-A option. Type-2-advisers, whose total share is given by β, split in two subgroups: Advisers in the first subgroup who represent share $\pi_c^*\beta$ of all advisers choose option A to uphold a positive image. Advisers in the second subgroup with population share $(1-\pi_c^*)\beta$ put their own payoff above image concerns and choose their preferred non-A options. The first sub-group then has mass only when they advisers assign diagnostic value to their choices, thus if $\gamma > 0$. Assuming that this is true, the following predictions can then be stated:

Prediction II. Suppose $\lambda > 0$. Then $\Pr[c = A|\text{bonus}] = \alpha + \pi_c^*\beta > \Pr[c = A|\text{no bonus}] = \alpha$

Second recommendation R2: As before, the dis-utility of inferring to be corruptible, thus to be a type-2-adviser, is given by $l_i > 0$. Since the advice in R2 is the same as in R1 we do not discount the diagnostic value by some $\lambda < 1$. The recommendation does not affect the adviser himself but the client. We thus assume, as for the first recommendation, that he suffers

[22] This also holds under the condition for weakly-increasing cdfs laid out in footnote 17 (it is the reason for the second expression in the *min*-term).

dis-utility from giving inappropriate advice, measured by k_i. Note that advisers initially formed a self-serving belief about $\tilde{u}$. In consequence, they have to stick to it. This means that there is additional dis-utility $k_i(1-\gamma)$, of not living up to one's prior motivated belief to $\tilde{c}^* = r_1$. An adviser's ex-ante utility from giving recommendation r_2, given his prior actions and beliefs, is then described by

$$V(r_2|r_2, r_1, c) = -k_i \left(\gamma \cdot \mathbb{1}[r_2 \neq c^*] + (1-\gamma) \cdot \mathbb{1}[r_2 \neq r_1]\right) - l_i \cdot \Pr[\theta = 2|r_2, c, r_1] \tag{8}$$

R2 – NO BONUS: Again, without a previous bonus type-2-advisers do not exist and $\Pr[\theta = 2|r_1, c] \leq \Pr[\theta = 2] = 0$. Since $r_1 = c = c^*$ was chosen initally, recommending $r_2 = c^*$ then maximizes (8). The share of advisers recommending option A (again) in NO BONUS is therefore α.

R2 – BONUS: For type-3-advisers, their previous behavior with $c = r_1 = c^* \neq A$ prevents them from inferring to be type-2-advisers since $\Pr[\theta = 2|r_2, c = r_1 \neq A] \leq \Pr[\theta = 2|c = r_1 \neq A] = 0$. Since for them $c = r_1 = c^*$ holds, they maximize (8) by recommending $r_2 = c = r_1 = c^* \neq A$.

First, consider the case that own actions in O had diagnostic value, thus $\lambda > 0$ and therefore $\pi_c^* \in (0, 1)$. Share $1 - \pi_c^*$ of type-2-advisers has then already revealed himself as such. For them, $\Pr[\theta = 2|r_2, c \neq r_1 = A] = \Pr[\theta = 2|c \neq r_1 = A] = 1$ applies. Their second recommendation r_2 is thus unaffected by image concerns. Accordingly, $r_2 = c^* \neq A$ maximizes (8) when $\gamma > \frac{1}{2}$ and $r_2 = r_1 = A$ when it holds that $\gamma \in (0, \frac{1}{2}]$.[23]

It follows that the mass of candidates for continued pooling with type-1-advisers in R2 is given by the overall share $\pi_c^* \beta > 0$ of advisers who has not yet revealed themselves to be type-2. They, together with type-1-advisers have a history of $c = r_1 = A$. By denoting the likelihood that a type-1-adviser chooses $r_2 = A$ with $\tau_{r_2} = \Pr[r_2 = A|\theta = 1]$ and the corresponding probability for a type-2-adviser who has not revealed himself by $\pi_{r_2} = \Pr[r_2 = A|\theta = 2, c = r_1 = A]$ we get the following posteriors:

$$\Pr[\theta = 2|r_2 = c = r_1 = A] = \frac{\pi_{r_2} \cdot \pi_c^* \beta}{\tau_{r_2} \cdot \alpha + \pi_{r_2} \cdot \pi_c^* \beta} \tag{9}$$

$$\Pr[t = 2|r_2 \neq A, c = r_1 = A] = \frac{(1 - \pi_{r_2}) \cdot \pi_c^* \beta}{(1 - \tau_{r_2}) \cdot \alpha + (1 - \pi_{r_2}) \cdot \pi_c^* \beta} \tag{10}$$

Analogously to the comparison of (5) and (6), (10) is larger than (9) whenever $\tau_{r_2} \geq \pi_{r_2}$. Repeating the analogous reasoning for an equilibrium with $\tau_c = 1$ in O, there is an equilibrium

[23] Note that when $\gamma > \frac{1}{2}$, their prior, self-serving belief leads even those advisers who have already revealed themselves to re-issue their biased advice for option A.

where all type-1-advisers choose option A for their second recommendation, thus with $\tau_{r_2} = 1$. This is the equilibrium on which we focus (see the above discussion on this selection for O, the same arguments carry over to R2).

Type 2-advisers who have so far not revealed themselves through inconsistent actions (i.e. $c \neq r_1 = A$) face again a trade-off: On the one hand, they could recommend their preferred choice $r_2 = c^* \neq A$ to prevent the cost γk_i of giving inappropriate advice, based on their personally preferred action. However, this would then reveal them to be type-2s and get them dis-utility l_i. In addition, they would give inappropriate advice based on their self-serving belief $\tilde{c}^* = A = r_1$ they formed in R1 which would now create costs of $(1-\gamma)k_i$ when they recommend $r_2 \neq r_1$. The alternative is to continue in recommending option A to pool with type-1-advisers and therefore uphold a positive self-image. By using (9), together with $\tau_{r_2} = 1$, a type-2-adviser then recommends $r_2 = c = r_1 = A \neq c^*$ if and only if

$$-k_i\gamma - l_i \cdot \frac{\pi_{r_2} \cdot \pi_c^* \beta}{\alpha + \pi_{r_2} \cdot \pi_c^* \beta} > -k_i(1-\gamma) - l_i \;\Leftrightarrow\; \frac{k_i}{l_i}(2\gamma - 1) < \frac{\alpha}{\alpha + \pi_{r_2}} \cdot \pi_c^* \tag{11}$$

In consequence, a type-2-adviser who re-issues biased advice by recommending $r_2 = A$ has low concerns of giving inappropriate advice (k_i) relative to their image concern (l_i). To formalize this, it will be useful to denote the family of cdfs of the ratio distribution k_i/l_i, conditional on an adviser's preferred option c^*, by $(R_{c^*})_{c^* \in \mathcal{S}}$. For example, a typical member is $R_B = \Pr[k_i/l_i \leq x | c^* = B]$.[24]

First consider the case that $\gamma > \frac{1}{2}$. Again, revealing one-self by recommending a non-A option increases the opportunity cost of doing so as pooling becomes easier. In equilibrium, advisers take this into account. From (11), it then follows that the share $\pi_{r_2}^*$ of hitherto not revealed type-2-advisers who continue to pool with type-1s has to solve the following expression:

$$\begin{aligned} \pi_{r_2} &= \sum_{c^* \in \mathcal{S} \setminus \{A\}} \alpha_{c^*} R_{c^*} \left(\frac{\alpha}{(2\gamma - 1)(\alpha + \pi_{r_2} \cdot \pi_c^* \beta)} \right) \\ &= \alpha_B R_B \left(\frac{\alpha}{(2\gamma - 1)(\alpha + \pi_{r_2} \cdot \pi_c^* \beta)} \right) + \alpha_C R_C \left(\frac{\alpha}{(2\gamma - 1)(\alpha + \pi_{r_2} \cdot \pi_c^* \beta)} \right) \end{aligned} \tag{12}$$

By analogous reasoning as for the RHS of (12), the above RHS is strictly less than one. It is also non-increasing in π_{r_2}. Therefore, there has to be a unique intersection $\pi_{r_2}^* \in (0,1)$ with the 45-degree line over the unit interval. We then get the following:

Prediction III.a) $\Pr[r_2 = A | \text{bonus}] = \alpha + \pi_{r_2}^* \pi_c^* \beta > \Pr[c = A | \text{no bonus}] = \alpha$ when $\gamma \in (\frac{1}{2}, 1]$.

Alternatively, if $\gamma \in (0, \frac{1}{2}]$ the second inequality in (11) is always fulfilled since its RHS is

[24] Since k_i and l_i are positively-valued and their distributions are commonly known, R_{c^*} is defined and also commonly known.

strictly positive while the LHS is strictly negative. It then follows that $\pi^*_{r_2} = 1$ and all of the unrevealed type-2s choose $r_2 = A$. In addition, the share $1 - \pi^*_c$ who have previously revealed themselves also choose $r_2 = A$ (see above). This prediction then follows:

Prediction III.b) $\Pr[r_2 = A|\text{bonus}] = \alpha + \beta > \Pr[c = A|\text{no bonus}] = \alpha$ when $\gamma \in (0, \frac{1}{2}]$.

Lastly, consider $\gamma = 0$. Own choices then have no diagnostic value as $\lambda = 0$. The main difference to the preceding analysis is that not choosing $c = r_1 = A$ for type-2-advisers does not necessarily reveal them to be of this type. In consequence, there is no mass $\pi_c\beta$ of candidates for *continued* pooling but *all* type-2-adviser are candidates for pooling with the moral type-1-advisers in R2 and none has previously revealed. The mass of those who potentially mimic type-1-advisers is thus given by β. The analogs to the inference posteriors (9) and (10) are equivalent to setting $\pi_c = 1$ in these expression.[25] Also, they are independent of the adviser's previous choice c since it does not have diagnostic value because $\lambda = \gamma = 0$ applies. Expression (11) then becomes

$$-l_i \cdot \frac{\pi_{r_2} \cdot \beta}{\alpha + \pi_{r_2} \cdot \beta} > -k_i - l_i \;\Leftrightarrow\; \frac{k_i}{l_i} > -\frac{\alpha}{\alpha + \pi_{r_2}} \tag{13}$$

and is always fulfilled, thus all type-2-advisers re-recommend $r_2 = A$:

Prediction III.c) $\Pr[r_2 = A|\text{bonus}] = \alpha + \beta > \Pr[c = A|\text{no bonus}] = \alpha$ when $\gamma = 0$.

From predictions III.a) through III.c) we get that for any weight $\gamma \in [0, 1]$, option A is more often re-recommended in BONUS than in NO BONUS, thus prediction 3 in the main text.

[25] Note that in slight contradiction to the initial definition of π_c as the share of type-2-advisers which behaves consistently in the own choice, setting this value equal to one does not mean that all behave consistently. It is however mathematically equivalent to this situation since the choice c has no diagnostic value. This is the same as if all type-2-advisers would have pooled with type-1-advisers. In both cases, the mass for (continued) pooling is the same and given by β.

Appendix B – Further data and analysis

	(1)	(2)	(3)	(4)
BONUS	0.440***	0.420***	0.441***	0.467***
	(0.047)	(0.041)	(0.059)	(0.060)
Personal Controls	no	yes	no	yes
Session Controls	no	no	yes	yes
Observations	99	99	89$^{\diamond}$	89$^{\diamond}$

Table 8: Average marginal effect of probit estimates for recommending option A in R1

	(1)	(2)	(3)	(4)
BONUS	0.170**	0.175**	0.210**	0.217***
	(0.074)	(0.071)	(0.090)	(0.071)
Personal Controls	no	yes	no	yes
Session Controls	no	no	yes	yes
Observations	99	99	79$^{\diamond}$	79$^{\diamond}$

Table 9: Average marginal effect of probit estimates for choosing option A for oneself in O

For the above tables:
Robust standard error in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Personal controls: age, gender, monthly budget, subject's region of origin and field of studies.
Observations with $^{\diamond}$: some combinations of the control variables predicted outcomes perfectly which is why the respective observations are not used in the ML-estimation.

	(1)	(2)	(3)	(4)
BONUS	0.194***	0.251***	0.182**	0.284***
	(0.071)	(0.072)	(0.087)	(0.107)
Personal Controls	no	yes	no	yes
Session Controls	no	no	yes	yes
Observations	99	87$^{\diamond}$	81$^{\diamond}$	66$^{\diamond}$

Table 10: Average marginal effect of probit estimates for recommending option A in R2

	(1)	(2)	(3)	(4)
BONUS	0.409**	0.508**	0.584**	0.624***
	(0.206)	(0.212)	(0.230)	(0.234)
Personal controls	no	yes	no	yes
Session controls	no	no	yes	yes
Observations	99	99	99	99

Table 11: Ordered probit estimates on the self-assessed preference for risk (Likert scale, 0 to 10)

For the above tables:
Robust standard error in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Personal controls: age, gender, monthly budget, subject's region of origin and field of studies.
Observations with $^{\diamond}$: some combinations of the control variables predicted outcomes perfectly which is why the respective observations are not used in the ML-estimation.

	NO BONUS		BONUS		OVERALL		rank-sum/χ^2-test
	mean	s.d.	mean	s.d.	mean	s.d.	p-value
age	24.824	8.002	23.208	5.411	24.040	6.882	0.264
male	0.451	0.070	0.354	0.070	0.404	0.050	0.339
region of origin							0.194
UK or Ireland	0.196	0.401	0.063	0.244	0.131	0.034	-
other Europe	0.137	0.348	0.188	0.394	0.162	0.370	-
N. America/Australia/New Zealand	0.020	0.140	0.083	0.279	0.051	0.220	-
South America	0.039	0.196	0.021	0.144	0.030	0.172	-
Asia	0.608	0.493	0.645	0.483	0.626	0.486	-
other	0.000	0.000	0.000	0.000	0.000	0.000	-
degree							0.220
bachelor	0.607	0.493	0.500	0.505	0.555	0.050	-
master	0.353	0.483	0.479	0.504	0.414	0.050	-
phd	0.000	0.000	0.000	0.000	0.000	0.000	-
other postgraduate	0.000	0.000	0.020	0.144	0.101	0.100	-
none	0.039	0.196	0.000	0.000	0.020	0.014	-
subject							0.261
economics/business/finance	0.216	0.415	0.375	0.489	0.293	0.457	-
other social sciences	0.353	0.483	0.229	0.425	0.293	0.458	-
psychology	0.059	0.237	0.021	0.144	0.040	0.198	-
public administration	0.039	0.196	0.062	0.244	0.051	0.220	-
math/sciences/engineering	0.157	0.367	0.083	0.279	0.121	0.328	-
arts or humanities	0.157	0.367	0.146	0.357	0.152	0.360	-
other	0.020	0.140	0.083	0.279	0.051	0.220	-
monthly budget (in GBP)	606.275	450.719	640.00	563.775	622.626	506.328	0.964
number of observations	51		48		99		

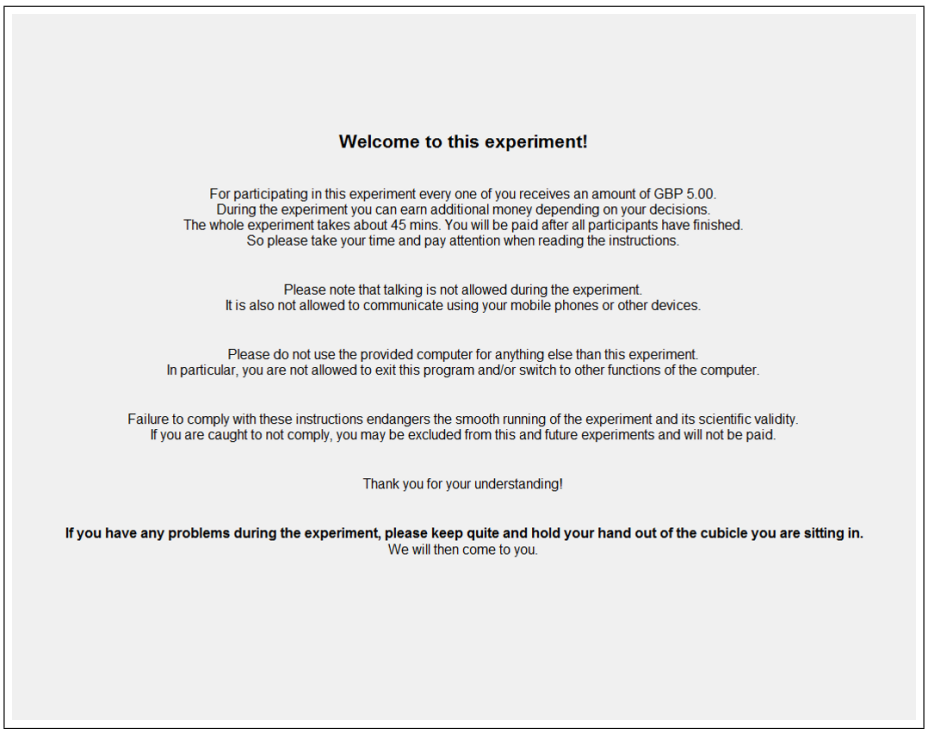
Table 12: Summary statistics for advisers' personal characteristics and dummy variable based on categorical data. The rightmost column provides p-values for a randomization check between NO BONUS and BONUS (Wilcoxon rank-sum tests for the variables age and budget; χ^2-tests for the remaining categorical variables).

Appendix C – Experimental instructions

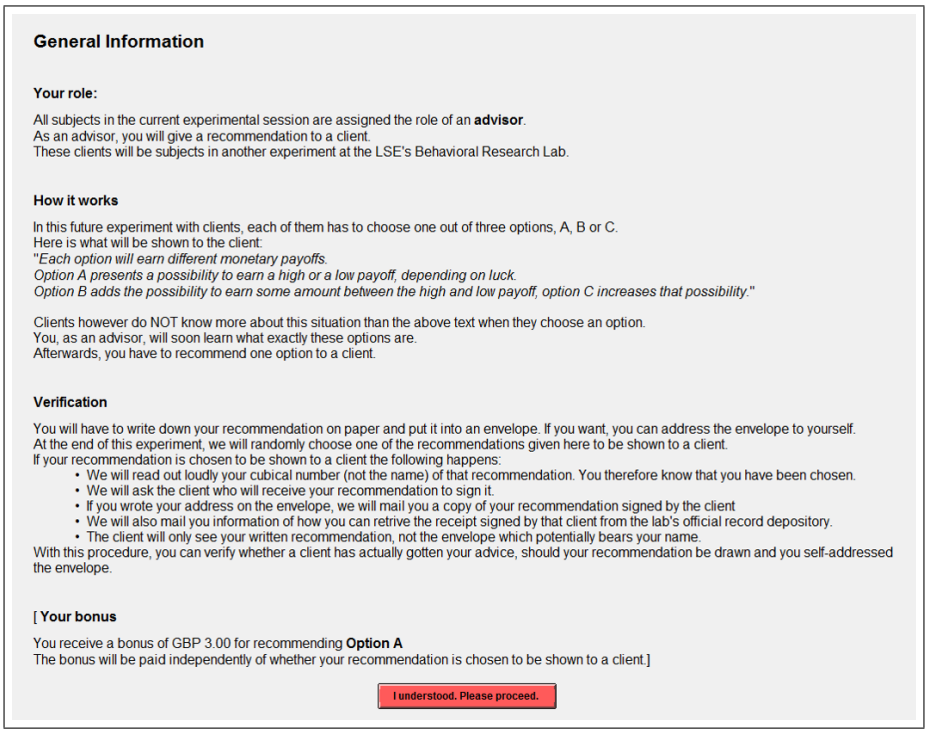
The following pages contain screenshots of instructions shown to subjects in ztree and on the information about the investment options printed on paper. They are presented in the order as they were seen by the subjects in the experiment.

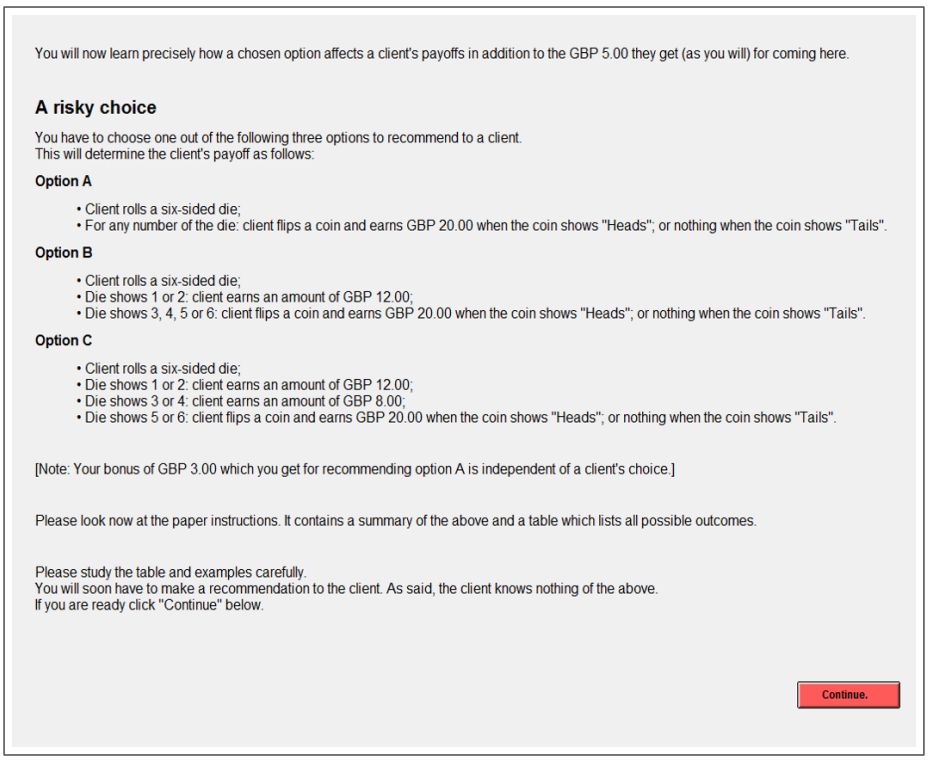
- Screen 1: Welcome stage and general instructions
- Screens 2a and 2b: Explanation for R1. Two screens which explain the client's choice situation, the adviser's role, and the investment options.
- Information on the investment options shown to advisers, printed on paper
- Screen 2c: Instructions for giving the first recommendation R1
- Screen 3: Instructions for making the own choice O
- Screen 4: Instructions for giving the second recommendation R2
- Screen 5: Exit questionnaire

The screens show the information shown to advisers in treatment BONUS. The parts which are not shown to advisers in NO BONUS are put in square brackets.



Screen 1





Screens 2a (top) and 2b (bottom)

A risky choice

One of the following options must be chosen. Then the following happens:

Option A:

- Roll die: for every outcome, play the lottery.

Option B:

- Roll die: if it shows 1 or 2, one earns GBP 12.00 for sure;
- Roll die: if it shows 3, 4, 5 or 6, one has to play the lottery

Option C: receive a chance to roll the same six-sided die:

- Roll die: if it shows 1 or 2, one earns GBP 12.00 for sure;
- Roll die: if it shows 3 or 4, one earns GBP 8.00 for sure;
- Roll die: if it shows 5 or 6, one has to play the lottery

The lottery:

For the lottery one has to toss a coin. “Heads” then yields GBP 20.00, “Tails” nothing.

Each row of the table below represents a possible result of the die. The columns describe the possible consequences, depending on the chosen option.

Die equal to….	**Option A** is chosen	**Option B** is chosen	**Option C** is chosen
1 or 2	lottery: GBP 20 or 0	GBP 12	GBP 12
3 or 4	lottery: GBP 20 or 0	lottery: GBP 20 or 0	GBP 8
5 or 6	lottery: GBP 20 or 0	lottery: GBP 20 or 0	lottery: GBP 20 or 0

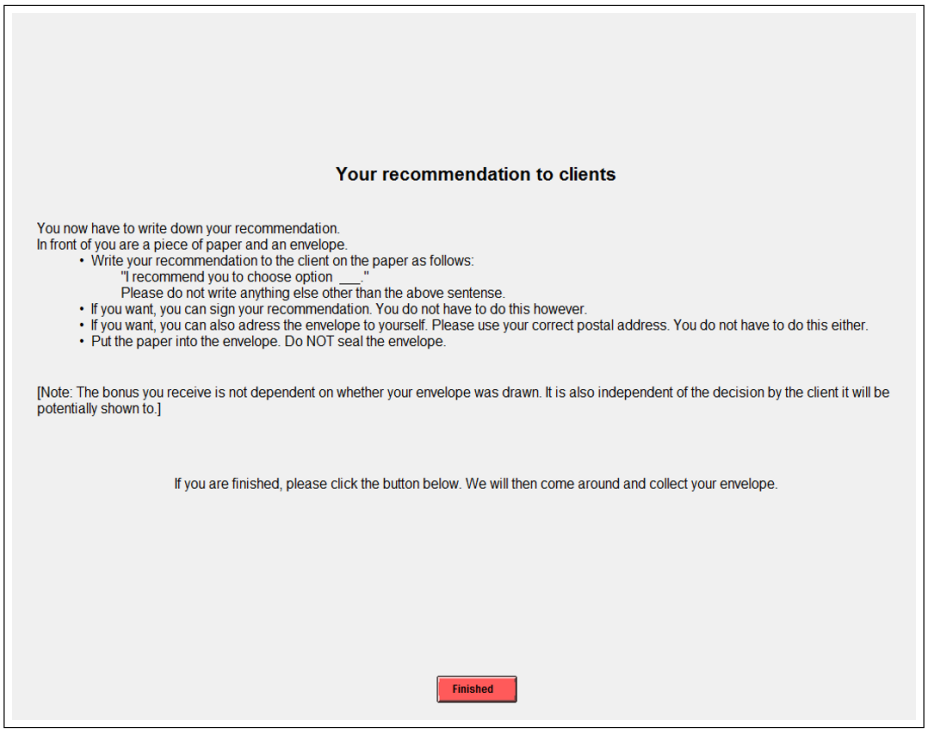
Example:

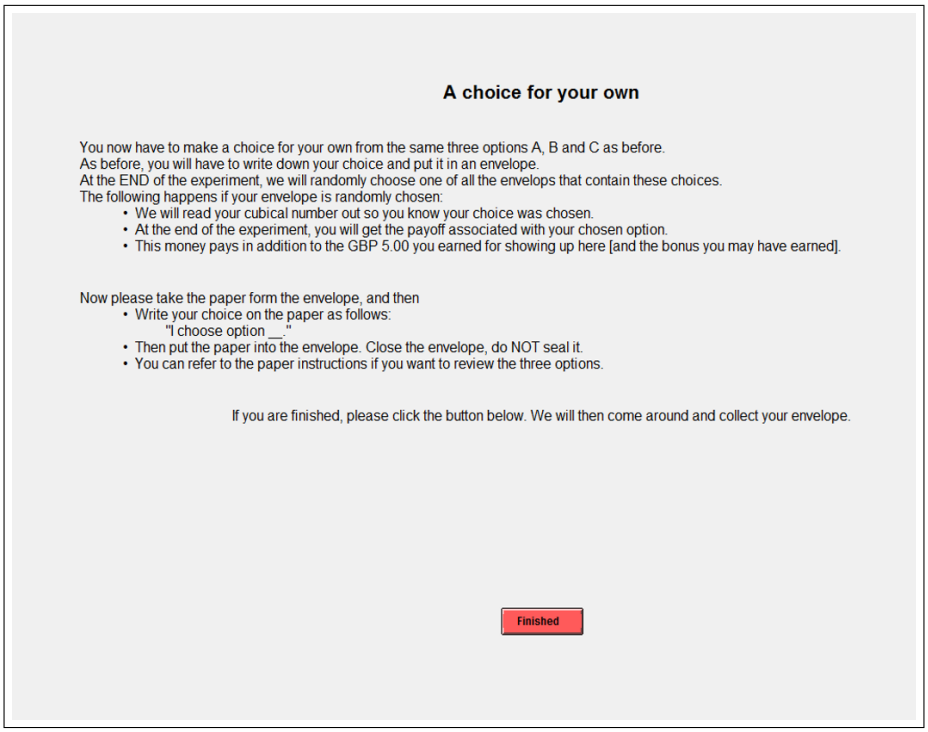
Suppose the die yielded 3: If option A or B was chosen before, one has to play the lottery. If option C was chosen, one would have gotten GBP 8.00 for sure instead.

Suppose the die yielded 1. If option B or C was chosen before, one gets GBP 12.00 for sure. If option A was chosen, one plays the lottery instead.

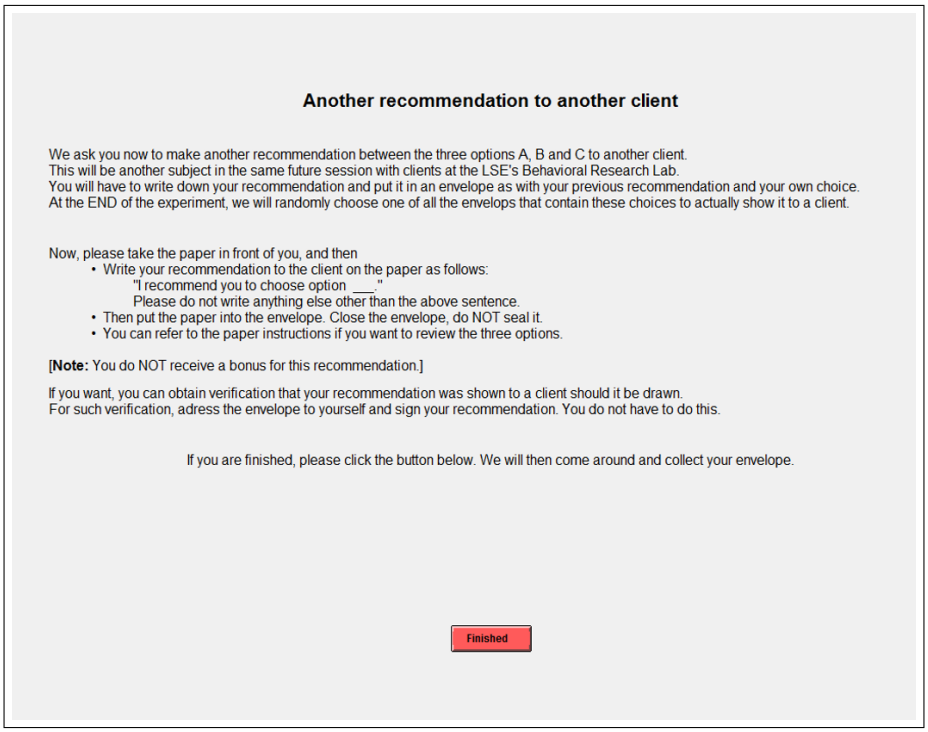
Suppose the die yielded 6. Independently of the chosen option one plays the lottery.

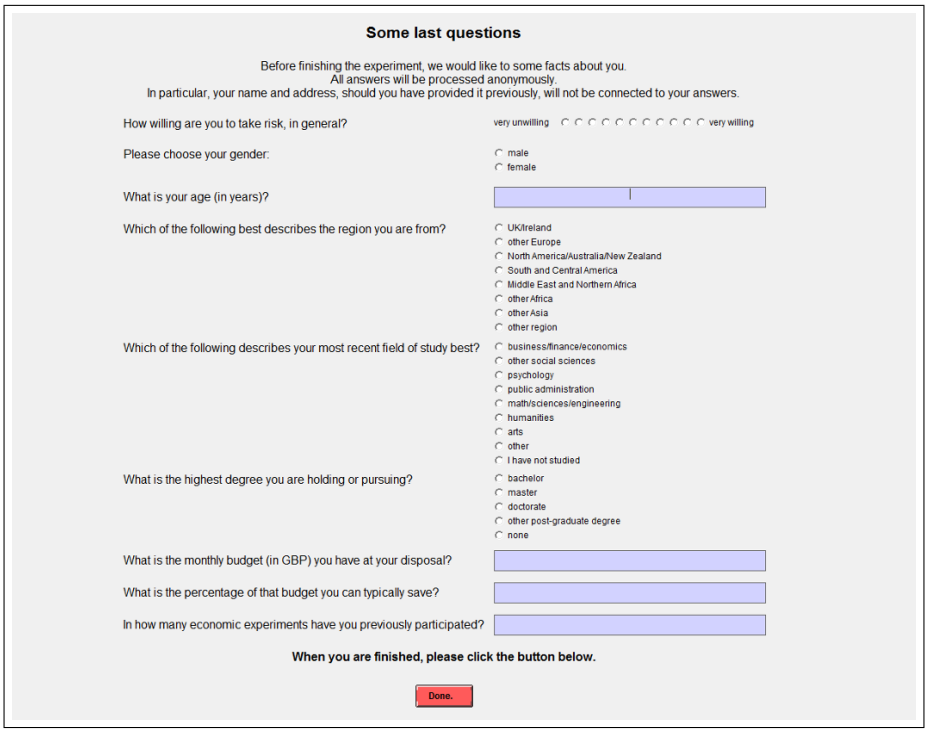
Information sheet shown to advisers
(It was placed face down on each adviser’s table with the following print on its back: "Information – do not turn until explicitly told so".)





Screens 2c (top) and 3 (bottom)





Screens 4 (top) and 5 (bottom)