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 (2017) : 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/10419/173410

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

Revised version, July 2017

Persistent bias in advice-giving

Zhuoqiong (Charlie) Chen
London School of Economics*

Tobias Gesche
University College London**

July 30, 2017

Abstract

We show that a one-off incentive to bias advice has persistent effects. In an experiment, some advisers were paid a bonus to recommend a lottery which only risk-seeking individuals should choose to a less informed client. Afterwards, they had to choose for themselves and make a second recommendation to another client, without any bonus. These advisers choose the risky lottery and recommend it a second time up to six times more often than advisers in a control group who were never offered a bonus. These results are consistent with a theory we present which is based on advisers' image concerns of appearing incorruptible.

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

** t.gesche@ucl.ac.uk, Department of Economics, University College London

We thank Björn Bartling, Antonio Cabrales, Lucas Coffman, Erik Eyster, Johannes Kunz, David Laibson, Ferdinand Langnickel, Ryan O. Murphy, Kristóf Madarász, David de Meza, Rani Spiegler, Roel van Veldhuizen, Joel van der Weele and several seminar and conference audiences for comments and feedback. Tobias Gesche also acknowledges support by the Swiss National Science Foundation (SNSF Grant #P2ZHP1_171900).

1 Introduction

Giving advice is at the heart of many professions where experts use their knowledge to guide less-informed clients on difficult and risky decisions. However, advisers often face a conflict of interest. Incentives such as sale commissions and kickbacks lead them to ignore clients' actual needs and advertise specific products, most prominently for financial advice (Mullainathan et al., 2012; Malmendier and Shanthikumar, 2014). Such conflicts can even be generated more subtly through unconditional gifts (Malmendier and Schmidt, 2017) – their consequences are however vast: For retirement investment in the US alone, which is just a share of the overall market for advised funds, conflicted advice is estimated to cause a 12% loss over returns for 30-year-savings. This corresponds to losses of 17 billion dollars per year (CEA, 2015). In other domains, for example when doctors advise patients on risky treatments, conflicted advice is also a problem (Dana and Loewenstein, 2003; Cain and Detsky, 2008) and stakes might even be higher, albeit harder to quantify. In light of these economic and ethical problems and the fact that disclosure often does not help, removing the cause of the conflict of interest seems appealing.[1] In fact, in many major economies policies which aim at removing and banning adverse incentives for advisers have been brought up or are being considered.[2] This paper explores how advisers respond to a conflict of interest and, more importantly, how they react when this conflict is removed.

In an experiment, we document that many advisers' recommendations are biased and that this bias persists, even after the underlying conflict of interest has disappeared. Subjects who acted as advisers were paid a bonus when they recommended a risky lottery which only a risk-seeking person would prefer. Clients received this advice and did not know the payoffs and the distribution for this lottery or its alternatives. Advisers had this information. About half of the advisers recommended the risky lottery when they were offered a bonus whereas only a small minority did so in a control condition without a bonus. We then removed the bonus and let advisers choose for themselves. Afterwards, they also had to issue a second recommendation to another client who had not received advice before. Advisers who were previously offered the bonus were more than thrice as likely to choose the risky lottery for themselves than those in the control condition. The removed bonus

[1] There is now numerous evidence that disclosing, as opposed to removing, conflicts of interest of experts may not only be ineffective, but also backfire, for example in Cain et al. (2005), Koch and Schmidt (2010) or Cain et al. (2011). Loewenstein et al. (2014) reviews the psychological literature and mechanisms underlying these effects; complementing economic accounts are presented in Li and Madarasz (2008), Inderst and Ottaviani (2012), and Gesche (2016).

[2] In the US, laws which would impose a fiduciary duty on retirement advisers are currently being discussed. Such a duty would prevent them from taking side payments which can affect their advice. In the UK, the "Retail-Distribution-Review" which bans commission-based financial advice came into force in 2013. Other European countries differ in the degree to which they regulate incentives which can lead to biased advice. Proposed policies range from banning them altogether (e.g. Netherlands), for some services (e.g. Italy) or not at all and just requiring disclosure (e.g. Germany). The MiFID II-directive by the European Union, which will become effective in 2018, also calls for avoidance of conflicts of interest in financial advice-giving.

also affected future recommendations and thereby the advisers' role as information multipliers: In the second recommendation, advisers who had previously been exposed to the bonus were six times more likely to recommend the risky lottery to another client as advisers in the control condition.

These findings are consistent with a simple theory we present. Its underlying reasoning is based on two main notions. The first is that being influenced by a bonus, i.e. giving biased advice, is deemed immoral. Consistent with this, almost half the advisers in our experiment who were initially offered the bonus did *not* recommend the risky lottery. The second notion is that people want to avoid the inference that their initial advice was biased. To signal one's own moral integrity, advice has to be unaffected by the bonus. This therefore requires consistency in advice-giving, even when this entails repeating biased advice after the bonus was removed. By similar reasoning, advisers may then even have to choose for themselves according to their previous, biased advice. In line with such a mechanism, we estimate that about 35 to 45% of those advisers whose initial advice was biased recommended the risky lottery again.

Our results also allow to explore how advisers determine appropriate advice. As for the second recommendation, 35 to 45% of the advisers whose initial advice was corrupted by the bonus also chose the risky lottery for themselves. This finding suggests that they linked their own choices to what they consider appropriate advice and then have to choose accordingly to avoid signaling their corruptibility. It speaks against an alternative reasoning where they could have self-servingly assumed that their clients are risk-seeking as then, there would not have been a negative signal if they had chosen for themselves a lottery different from their previous recommendation. Also, had these findings not been established in a controlled experiment, they would be in line with risk-seeking advisers self-selecting into compensation schemes which reward advertising risky choices. However, the fact that we exogenously vary the presence of the bonus allows us to conclude that the bonus, in a persistent and causal way, affected advisers' repeated recommendations and their own choices.

Related literature: Recent findings on adviser behavior resonate strongly with our results: Foerster et al. (2016) use observational data on about six thousand Canadian financial advisers and more than 580,000 of their clients. They show that not clients' personal characteristics but simple fixed effects for the individual advisers explain most of the variation in how risky the clients' investment portfolios are. Using the same data set, Linnainmaa et al. (2016) document that recommendations to clients are also reflected by the choices which advisers make for themselves: The advisers in their sample chose the same return-chasing and actively managed funds which their respective clients held. They also show that advisers did so, even though these investments were riskier than others and, net of fees, performed worse than an index fund would have performed. Our findings show how commissions (which are typically paid for selling such funds) can explain these patterns.

Closely related to our work is an experiment in Gneezy et al. (2016). In it, advisers also faced a bonus to recommend a specific lottery to uninformed clients. The authors then show that for this one-shot recommendation, there is a relatively large bias when advisers first learned about the bonus before seeing the lotteries. This bias decreases when they first see the lotteries and then learn about the bonus. A similar pattern is also observed in Babcock et al. (1995): In their experiment subjects acted in fictitious roles of plaintiff or defendant in a legal case. They report that these subjects find it much harder to agree on a settlement value when they knew their role before learning about the case's details than vice versa. In the same experimental paradigm, Loewenstein et al. (1993) show that subjects in the hypothetical role of plaintiff or defendant shifted what they considered a fair settlement value towards their respective positions.

These findings resonate with a more fundamental mechanism in which role-induced dispositions lead people to align their judgment such that it serves their role. Konow (2000) finds a systematic shift for entitlements to the contributions of a collectively generated surplus. This shift in entitlement depends on the roles the subjects had in a subsequent dictator game in which this surplus was split. A classic study which reports a similar effect is Festinger and Carlsmith (1959). In it, the authors demonstrate that paying subjects to report favorably about an unpleasant task to others improves their subsequent evaluation of this task, relative to when they were not paid. They explain this by peoples' inherent desire to minimize cognitive dissonance, which would arise in this context by having advertised the unpleasant task (for money) but not expressing an aligned personal opinion. That such induced judgments carry over to own actions in a harmful way, i.e. that people self-deceive (Trivers, 2011), is demonstrated in a recent study by Schwardmann and van der Weele (2016). They show that when tasked to convince others of their own ability, subjects overstate their own ability in subsequent private self-assessments, even when doing so is costly to themselves. Mijović-Prelec and Prelec (2010) report a similar pattern of costly self-deception and how it can be derived as a consequence of self-image concerns.[3]

Our work takes up on these insights and demonstrates how these effects diminish the effectiveness of removing conflicts of interest in advice-giving. In the theory we lay out, consistency with biased advice in own choices and repeated advice is instrumental in preserving a positive image. A similar logic underlies the finding by Falk and Zimmermann (2017a,b). They show that subjects forfeit opportunities to improve their accuracy in order to be consistent in an estimation task and thereby signal ability to a principal or themselves. Eyster (2002), in a related theoretical account, also explores similar implications of consistency as a means to disguise past mistakes. We show how

[3]For a further discussion about how cognitive dissonance and (motivated) self-perception relate see Kunda (1992). For models which incorporate Festinger (1957)'s concept of cognitive dissonance into a formal economic framework, see Akerlof and Dickens (1982) and Rabin (1994).

such harmful consistency can emerge through conflicts of interest and how it persistently affects the role of advisers as information multipliers. Therefore, we also connect to the recent literature on the adverse effects of bonus payments (Christoffersen et al., 2013; Bénabou and Tirole, 2016), in particular for financial services and how the resulting conflicts of interest shape the self-perception and attitudes of those exposed to these incentives (Zingales, 2015; Cohn et al., 2014). However, our findings come from a neutral framing and relate to biased advice more generally.

The findings we present and the theory we develop also connect to the literature on moral reasoning and economic behavior. Central to these approaches is the notion that people care about their image of being perceived as moral and that their own actions are used to infer about themselves (Bodner and Prelec, 2003; Bénabou and Tirole, 2004), in particular, their own moral values (Bénabou and Tirole, 2011; Falk and Tirole, 2016). Although image concerns can refer to both, social and self-image, the latter alone can steer moral behavior. This applies, for example, to non-maximal lying in order to uphold the illusion of being honest (Mazar et al., 2008), inflicting less harm on others while seeing oneself on a video screen (Falk, 2017), or paying more under a pay-what-you-want scheme than under fixed prices to avoid appearing greedy to oneself (Gneezy et al., 2012). The theory we present and the results we report show how image concerns can be linked to moral reasoning in the context of advice-giving and how they lead to persistent biases in a setting where social image concerns were minimized.

We also shed more light on how people process information which threatens their self-image. In a recent review on the topic, Gino et al. (2017) summarize findings from other domains which demonstrate that often, information is not processed in an objective manner. Rather, people often act as "motivated Bayesians" who employ uncertainty and ambiguity in a self-serving way (Dana et al., 2007; Haisley and Weber, 2010; Exley, 2016). In particular, this includes the formation of beliefs about others and others' preferences to accommodate own selfish actions (Di Tella and Pérez-Truglia, 2015). Our results indicate a link between advice and own choices, rather than on self-serving beliefs about others. This connects to earlier research which shows that people base their estimates of others' preferences, particularly risk preferences, on their own (Mullen et al., 1985; Faro and Rottenstreich, 2006).

The next section describes a mechanism of how moral and self-image concerns can lead to persistent bias after advisers had a conflict of interest. Section 3 explains the design and procedures of our experiment for which we derive predictions, based on the behavioral mechanism we propose in Section 4. Section 5 presents our results. Section 6 discusses our findings and concludes by reviewing their implications for the economics of 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 instructions.

2 A psychological mechanism

In this section, we describe how a preference to appear as an unbiased adviser can lead to a repeated bias in further advice and in own choices. The underlying mechanism can also be derived in a formal model (see Appendix A). It is based on an adviser ("he") who advises a client ("she") and who is concerned with what his current actions reveal about his past motivations to give advice. Specifically, it assumes that an adviser's overall utility consists of three elements: 1) utility derived from expected monetary payoffs, 2) psychological or material costs of not giving appropriate advice, and 3) diagnostic dis-utility of learning from one's current actions that previous advice was biased.

While the first element is standard, the second reflects advisers' uneasiness to recommend something they do not consider advisable to a client. For example, an adviser might think that a recommendation for a particular choice is suited for his client because, given the adviser's belief about the client's preferences, this would be the client's preferred choice if she had the same information as the adviser. Not recommending this preferred choice then creates costs because one has not acted in the client's best interest.[4] 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 trained financial advisers who receive information about the clients' characteristics and who have no conflict of interest have difficulties in predicting their client's risk preferences (Roth and Voskort, 2014). When there is a conflict of interest, this uncertainty can be instrumentalised in a self-serving manner: Advisers may form a self-serving belief about their clients' preferences such that this belief is compatible with their, potentially biased, advice.

There are, however, 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).[5] Some advisers might also determine what they consider an advisable action by the answer to the question "What would I choose if I were in the client's position?". In consequence, advisers' own preferences can also play a role in determining the choice of which they think that it should ideally be recommended. What advisers consider as advisable actions can thus be independent of their own preferences (when the former are formed based on a self-serving belief) or the same. Our experiment allows to explore the different implications which each of these different lines of reasoning has. However, for the principal

[4] In fact, for many adviser-client-relations such as doctors and patients, lawyers and clients, and several situations of financial advice-giving, there is a fiduciary duty which legally requires the adviser to act in the client's best interest.

[5] Even though initially coined by Ross et al. (1977) as a "false consensus effect", the falsity of estimating others' preferences based on one's own is not evident. Works by Hoch (1987) and Dawes (1990) demonstrate that often, 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 preferences. Engelmann and Strobel (2000) show that subjects do so when they are incentivised to make accurate predictions.

mechanism of repeated biased advice we explore, it only matters that costs of not recommending an advisable choice exist, independent of how this advisable choice is determined.

The third factor which matters for advisers is diagnostic (dis-)utility they derive when they learn to have given biased advice, based on a model of self-signaling. In contrast to the costs of giving inappropriate advice, 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 adviser's motive for giving advice, e.g. whether prior advice was biased or not. The important implication of such an inference is that advisers can only uphold a positive image of themselves as long as they do not take actions which are incompatible with this notion. Dual-self models have been used previously to explore how people infer about themselves, in particular, their morality (e.g. Bodner and Prelec, 2003; Bénabou and Tirole, 2011; Grossman and van der Weele, 2017). Here, we use such an approach 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 conflicts of interest affect advisers' choices and their recommendations. To see this, consider an adviser whose recommendation is corrupted: His cost of giving biased advice, i.e. the cost of recommending something he does not consider advisable, is thus smaller than the (material) benefit he gets from giving such advice. If the adviser also is sufficiently concerned about his self-image, he then needs to continue to give the same biased advice again when the conflict of interest has disappeared. The reason is that in order to entertain the notion that the initial advice was unbiased, it should be unaffected by the presence of any external incentive. However, changing advice after the conflict of interest disappeared signals the opposite.

As mentioned above, an adviser's own preference can also determine what he thinks should be recommended to a client. In this case, an adviser whose advice was corrupted and who has then to choose for himself is put on the spot. If he does not choose as he has recommended, he also signals his previous corruptibility. Accordingly, if their image concerns are sufficiently high, advisers effectively bias their own choices. In summary, a behavioral trait which generally seems to be desirable – a preference to be perceived as unbiased – can lead to a persistent bias in the context of advice-giving. In addition, it can have a lasting effect on advisers' own choices. The following experimental design then allows to empirically explore, in a controlled environment, such persistent bias in repeated advice-giving and in own choices.

[6] In the context of this paper, we mainly talk about image concerns as self-image concerns because this is, as we will argue, most consistent with our experimental setup. However, the main mechanism we propose and which underlies a persistent bias can also be triggered by social image concerns, i.e. when an outside observer's perception matters, or by social and self-image concerns together.

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 the screen. Subjects who acted as advisers were informed that they would get GBP 5.00 as a show-up fee for participating in the experiment (GBP 1 = USD 1.43 at the time of the experiment) and that there would be further possibilities to earn money. They were then informed that they would act as advisers for clients in a future experimental session and that these clients would be drawn from the same pool of subjects and receive the same show-up fee.

Advisers had to recommend one of three risky choices, referred to as option A, B, and C to their clients. They were informed that option A's payoff would be either high or low, depending on luck. Option B would also allow to get an intermediate payoff and option C increases the probability for getting this intermediate payoff. Advisers knew that this information would also be given to the clients but that clients would neither know the options' payoffs nor the respective probabilities. They, as advisers, would soon learn these parameters 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 computer interface see Appendix D). The text on the sheet explained the following procedure of how an option's payment 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 payment or a lottery. This lottery was described as a (fair) coin toss with heads yielding GBP 20 and tails nothing. Table 1 summarizes this. It was also printed on the paper instructions, together with a text and examples which explained this procedure in detail.

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 fair coin toss.

Throughout the experiment, advisers could keep this paper and refer back to it when needed. Note that a choice among the options, i.e. the three compound lotteries, allows to categorize the underlying risk preferences: If one compares option A and B, only a person who is willing to give up a safe payment of 12 to play a lottery with an expected payment of 10, i.e. a risk-seeking individual, chooses option A. Conversely, option C is preferred to option B only by a person who wants to sacrifice an expected payment of 10 for a safe payment of 8. Thus, only a risk-averse individual should choose option C. Accordingly, option B is chosen by a person who is neither sufficiently

risk-averse nor sufficiently risk-seeking. Reflecting this ordering based on risk-preferences we will henceforth refer to option $A/B/C$ as the risky/neutral/safe option, respectively.[7]

After these initial explanations, the experiment proceeded along five steps:

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 on a piece of paper that they recommend their client to choose either option A, B or C. They were then instructed to put this paper, which had the adviser's cubicle number on it, into an envelope and close it. The envelope was then collected by an experimenter and put into a box. Before they made their recommendations, advisers were told that at the end of the experiment, one of the envelopes would be randomly drawn from the box to be shown to a client who would then have to choose an option.

Step 2 – Own choice O: When all advisers had written down their recommendation R1 and all envelopes were collected, they were informed that they would now have to choose one of the three options for themselves. Advisers were previously not informed about this step. The procedure was the same as for issuing advice: Subjects had to write their choice on a letter and 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. The corresponding adviser would then play the lottery implied by this choice. Ex-ante, the choice situation and its implementation probability were thus the same as the one for which they had previously issued advice in R1, except that the choice was for themselves.

Step 3 – Second recommendation R2: After advisers had made their own choice O, they were asked to make a second recommendation. Again, this was not announced beforehand. The procedure was exactly the same as for R1, including the collection of envelopes in a separate box and announcing in advance that one would be sampled from it. Advisers were also informed that their second advice, if it was sampled, would be shown to another client who would not have received any previous advice. The decision situation thus mirrored exactly the first advice in R1.

Step 4 – Questionnaire: When all recommendations from R2 were collected, advisers had to fill out a short on-screen questionnaire which elicited personal information. It also included a short question on advisers' general willingness to take risk.

[7] This choice between possible sub-lotteries within a compound lottery is essentially a stripped-down version of 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 pairs of lotteries. Across the ten choices, each pair's second lottery becomes increasingly risky. One of the ten choices is then randomly picked to be implemented. This 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 allow the categorization along risk-seeking/neutral/averse preferences.

Step 5 – Payoff: At the end of the experiment one envelope was sampled from each of the boxes for R1, O, and R2. The corresponding cubicle number (but neither the recommendation nor the choice by the subject) was read aloud so that the respective subject knew whether his or her envelope was sampled. The sampling at the end of the experiment was announced and explained before advisers made their respective recommendations and choices. Subjects were then called, one by one, to the laboratory's exit where they were paid privately. In each session, the subject whose own choice was chosen to be implemented also rolled a die and, if necessary, also tossed a coin to determine the chosen lottery's payoff. The subject was then paid accordingly.

NO BONUS versus BONUS treatment: The above describes the experimental procedure in our baseline condition to which we will refer to as NO BONUS. Our experimental manipulation was to offer some advisers a bonus for recommending the risky option A in R1. We will refer to this treatment as BONUS. After having been informed that they had to give advice but before seeing the sheet with the detailed information about the investment options, every second adviser in a given session was randomly chosen to be informed that they would get a bonus of GBP 3 if they recommend option A. This bonus was only paid for subject's first recommendation R1, together with other earnings at the end of the experiment. For the advisers in BONUS, it was clearly stated on screens which explained the O and R2 tasks that there would not be any additional bonus for choosing or re-recommending option A in these tasks.[8] This within-session, across-subjects intervention with regard to the bonus is the only difference between the NO BONUS and the BONUS-treatment.

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 used the following procedure: We allowed advisers to voluntarily sign their recommendations and address the envelopes to themselves. Advisers were explained that if their recommendation was chosen to be shown to 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 option before they made their first recommendation and reminded of it before the second. It was also announced that the sampled recommendation's cubicle number (but not the recommendation itself) would be announced at the end of this experiment. In this way, advisers

[8]Since advisers' payoffs 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. Clients were informed of the bonus when they received a recommendation R1 from an adviser who had been in the BONUS treatment.

knew that experimenters were pre-committed to actually show the sampled advice letters to the 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 (USD 9.55 at the time of the experiment) while no session lasted longer than 45 minutes. All subjects were students across several degrees and fields of studies. Table 11 in Appendix C 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. Before, the principal study design and research questions were submitted to the school's research ethics committee and its approval was obtained. The experimental interface was implemented using zTree (Fischbacher, 2007). A week after the eight adviser sessions, we invited sixteen additional subjects from the same pool for an additional session. In this session, they acted as clients and each received one of the sampled recommendations from the previous adviser sessions (eight for R1 and eight for R2). After reading the recommendation, clients made their choices and were paid the outcome of the independent lottery they chose. As mentioned above, advisers knew about this structure, in particular that their own and the clients' payoffs realizations, conditional on a given choice, would be identical and independent. In this paper, we focus on advisers and their actions.[9]

4 Predictions

In this section, we derive predictions for our experiment. They are based on the assumptions described in Section 2, thus on advisers maximizing their overall utility from direct pecuniary payoffs, costs of giving inappropriate advice, and image concerns of being perceived of having issued biased advice. Given our treatment intervention, we make the predictions with regards to how often the risky option A is recommended and chosen. All these predictions can also be derived from a formal model which can be found in Appendix A.

First recommendation R1: In NO BONUS, there is no pecuniary gain of issuing any specific recommendation. Since this is the first action an adviser takes, it does not have signaling value with regards to past behavior. Therefore, image concerns do not matter. Absent other motives, only the costs of issuing inappropriate advice remain. Thus, only advisers who think that option A is actually a good recommendation recommend it.

[9] Given that we have four relevant conditions (recommendation from R1 vs. R2 and BONUS vs. NO BONUS) and only sixteen client observations which are, due to our random sampling procedure, not balanced across these conditions, there is not much analysis which can be done due to limited statistical power.

In the BONUS treatment, advisers are paid for recommending option A. In addition to those who actually think that this option is advisable, some advisers might be induced to recommend it in order to earn the bonus although they do not consider option A advisable. This happens when the costs of giving inappropriate advice are low, relative to the pecuniary utility associated with the bonus. Assuming that this is true for some advisers, the following prediction can then be stated:

Prediction 1: *More advisers in BONUS than in NO BONUS recommend option A.*

Own Choice O: Different to the first recommendation R1, advisers now make a choice for themselves. Therefore, costs of giving inappropriate advice play no role. In NO BONUS, there are also no image concerns of having issued biased advice before. Accordingly, advisers in this condition choose the option which they prefer (which might coincide with what they consider advisable and thus have previously recommended, see below).

In the BONUS treatment this reasoning does not go through. As previous advice in R1 could have been corrupted by the bonus, there is also a concern of being perceived or perceiving oneself as a biased adviser. This is relevant when advisable choices relate to one's own preferences, for example, when the advisable choice is the answer to the question "What would I do if I were in the client's situation". The consequence of such reasoning is that for an unbiased adviser, his own choice in O and his previous recommendation in R1 should coincide. In the presence of image concerns, this has further implications for advisers who have initially recommended option A in R1 because they wanted the bonus. When they now choose differently they signal that their initial advice was corrupted. Biased advisers can avoid this negative signal if they also choose option A for themselves. This however leads to a loss in expected pecuniary utility as they choose option A instead of their truly preferred non-A choice. They therefore do so if the image costs are high, relative to this loss.[10] In addition, those advisers who would have chosen option A anyhow because they truly prefer it, as in NO BONUS, choose A for themselves. Assuming that there is such reasoning which relates own choices to appropriate advice and that image costs are high enough for some advisers, we get the following prediction:

Prediction 2: *More advisers in BONUS than in NO BONUS choose option A for themselves.*

If the above prediction were wrong, this could have two reasons. The first is that advisers determine advisable choices independently of what their own preferred choices are. For example, advisers whose advice in R1 was affected by the bonus might have formed a self-serving belief about the client's risk preference. This would allow them to rationalize their first recommendation for option A without incurring the costs of giving inappropriate advice. Such a self-serving belief about

[10] In Appendix A we show that there is no "reverted" signaling equilibrium in which, just due to image concerns, not the corrupted but the previously unbiased advisers who really prefer option A choose something different than this option for their own choice O.

the client's preferences does not need to relate to their own preferences. In this case, choosing differently for themselves and for their client would not send a signal that the previous advice was biased. The second reason why the above prediction might be wrong is simply that advisers do not have sufficiently high image concerns. Absent such concerns, the initial, biased recommendation should not affect the second recommendation. However, image concerns would require consistency between the first and the second recommendation to which we turn now.

Second recommendation R2: An adviser's own pecuniary utility is unaffected by his second recommendation. In NO BONUS, there was no incentive to bias advice and image concerns of having given biased advice do therefore not play any role either. Accordingly, only the costs of giving inappropriate advice matter. Option A is then only recommended by those who have already recommended it previously in R1 because they genuinely consider this option advisable.

In the BONUS treatment, the second recommendation does not affect the adviser's pecuniary payoff anymore because the bonus has been removed. However, the previous recommendation might have been biased through the bonus so that image concerns matter. A previously unbiased adviser should just recommend what he actually considers advisable and therefore, should repeat his initial advice. In order not to be perceived as biased, this means that advisers who have previously been corrupted by the bonus have to re-issue the same advice. As these advisers do not consider option A advisable, they have to bear the costs of giving inappropriate advice. When these costs are small, relative to their image costs, they thus mimic the behavior of incorruptible advisers who consider option A advisable by also recommending this option. In addition, these incorruptible advisers, as in NO BONUS, also re-recommend option A.[11] We thus get the following prediction:

Prediction 3: *More advisers in BONUS than in NO BONUS recommend option A.*

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 answer two main questions. First, by testing Prediction 3 we can find evidence in line with image concerns which cause repeated bias in advice-giving. Second, if this is true, Prediction 2 enables us to investigate this effect in more detail. Its confirmation would suggest that advisers' own choices and their biased recommendations are related. When they are not related, e.g. through self-serving beliefs about the clients' preferences, we would not expect this prediction to be confirmed.

Before we turn to our results, it is worth to note that the recent findings by Gneezy et al. (2016) are also consistent with the mechanism which leads to the above predictions. In fact, our experimental design for the first recommendation R1 in the BONUS-treatment is inspired by one of the experiments they report. In its "before"-condition, advisers were first shown two investment

[11] As for the own choice O we show in Appendix A that within our model, there is no "reverted" signaling equilibrium in which incorruptible advisers' actions are affected by image concerns.

options, call them option a and b.[12] Before advisers got further information on the options, they were informed that they would get a bonus if they recommended option a (which had a lower expected value but less variance than its alternative, option b). Advisers were then explicitly asked to consider which one they would recommend to a client. Afterwards, advisers had to give an actual recommendation to a client.[13] In their "after"-condition, advisers learned about the bonus after having initially considered which option to recommend but before actually recommending it. Gneezy et al. find that the bias towards option a is larger in the "before" than in the "after"-condition.

Their finding is, in principle, consistent with the mechanism we propose. To see this, think of the initial consideration as an initial recommendation. Consider an adviser in their "after"-condition who has initially considered option b as advisable and then learns about the bonus. If he then recommends option a to earn the bonus, this creates a very clear signal about his corruptibility, similar to a change in the first and second recommendation in our experiment. In contrast, advisers in the "before"-condition who know about the bonus and want to recommend option a can adjust their initial consideration accordingly. This prevents signaling their corruptibility so that image costs of recommending option a are lower than in the "after"-condition. However, if option a is a truly bad choice, such an advance adjustment of the initial consideration is harder to perform because there are hardly any incorruptible advisers who consider such an option advisable and whom one can mimic. In line with this, the authors also show in a further experiment that the before-after-difference disappears when option a is made a clearly inferior choice, i.e. if it is made first-order stochastically dominated by option b.

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 advisers to follow this incentive and recommend it more frequently than in NO BONUS. In fact, only 3.9% of advisers in NO BONUS recommended option A in their first recommendation whereas more than half of all advisers, 54.2%, in the BONUS treatment recommended this option. This increase by 50.3 percentage points is highly significant (Fisher exact test, two-sided: $p = 0.000$).[14] Figure 1 shows the overall distribution of the initial recommendations across treatments.

[12] In their experiment, these options were also called option "A" and "B", respectively, as in ours but had different parameters. To avoid confusion, we use lowercase letters to refer to their lotteries.

[13] They asked the advisers "*Please take a minute to decide which product to recommend*" [p.46] before advisers had to click a button stating that they were ready to issue their recommendation.

[14] Although we have directed hypotheses, p-values reported here and after refer to more conservative two-sided tests.

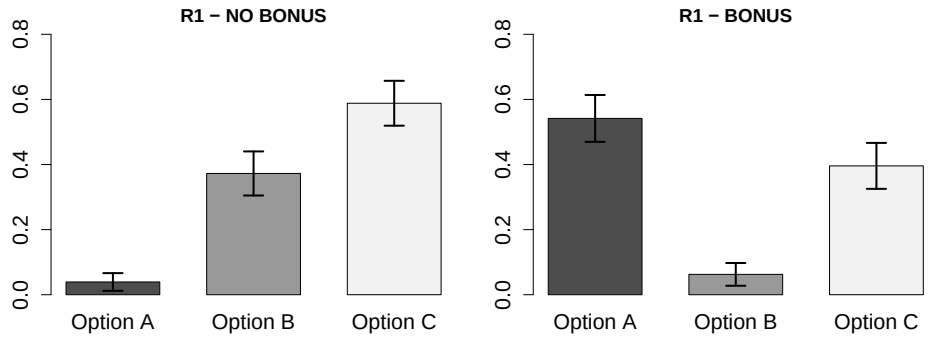


Figure 1: Frequency for each option being recommended in R1, together with standard errors.

We also employ a parametric approach by estimating the following regression model which includes additional control variables:

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

In the above, the dependent variable is an indicator which takes a value of one if a subject's first recommendation $r_{1,i}$ was for option A. $Bonus_i$ is a dummy which indicates 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 available budget, region of origin, the highest degree the subject holds or pursues and the field of studies. Session fixed effects are collected in $\mathbf{s}_i$. The error term ϵ_i captures idiosyncratic noise in an adviser's recommendation. Table 2 presents the OLS results when controls are successively added. It shows that the increase of about 50 percentage points in the probability of recommending option A through the bonus is almost unaffected by the addition of these controls and remains significant. This also applies if one looks at a probit model. Appendix C collects these probit results for each of the relevant linear models in the main text; the corresponding results are always very similar. Therefore, our results show that our treatment manipulation is effective and are in line with Prediction 1.

	(1)	(2)	(3)
$Bonus$	0.489***	0.501***	0.481***
	(0.093)	(0.076)	(0.092)
Personal Controls	yes	no	yes
Session Controls	no	yes	yes
Observations	99	99	99
Adjusted R^2	0.280	0.329	0.310

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

Note that our result also has an important second implication. Almost half of the advisers (45.8%) did *not* recommend option A, even when they were offered money to do so. This is

consistent with the notion that there exist non-pecuniary costs of giving such advice and that for a considerable fraction of advisers, these costs outweighed the utility of the bonus payment.

Results for O: For their own choice, no bonus was paid to advisers in both conditions. Figure 2 displays their choices. In the baseline NO BONUS, we observe that 9.8% chose option A for

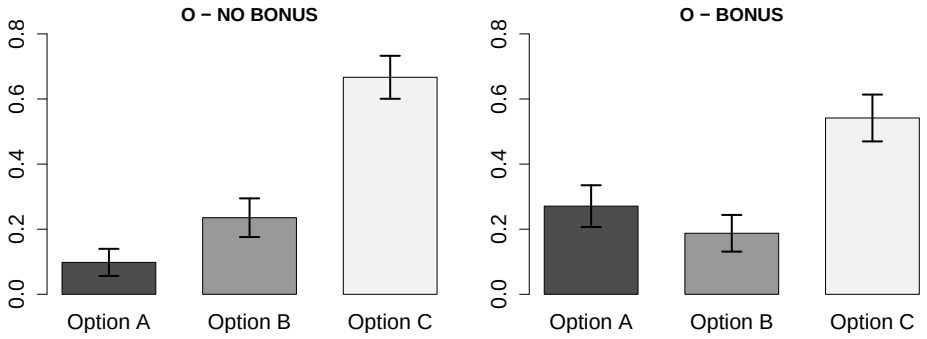


Figure 2: Frequency for each option being chosen in O, together with standard errors.

themselves. This share is comparable to the results of Holt and Laury (2002) who find that six to eight percent of subjects exhibit risk-seeking preferences. In BONUS, when advisers were *previously* offered the bonus for their first recommendation, the share of those who chose option A is 27.1%, almost three times as many advisers as in NO BONUS. This 17.3 percentage points increase is statistically significant (Fisher exact test, two-sided: $p = 0.036$). This finding is also confirmed by regression analysis. For this, we replace the dependent variable in model (1) with a dummy indicating whether an adviser chooses option A for himself. Columns 1 through 3 in table 3 report the corresponding results when the same control variables as in the preceding regression analysis are successively added. We therefore regard Prediction 2 as supported by our results.

	(1)	(2)	(3)	(4)	(5)
$Bonus$	0.219**	0.174**	0.218**	0.031	
	(0.095)	(0.078)	(0.087)	(0.087)	
$r_1 = A$				0.387***	
				(0.115)	
$\widehat{r_1 = A}$ (via $Bonus$)					0.453***
					(0.143)
Personal Controls	yes	no	yes	yes	yes
Session Controls	no	yes	yes	yes	yes
Observations	99	99	99	99	99
Adjusted R^2	0.065	0.019	0.088	0.219	0.225

*Table 3: OLS and 2SLS estimates of the probability that advisers choose option A for themselves in O. Robust standard errors in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Personal controls: age, gender, monthly budget, subject's region of origin, degree and field of studies.*

Given these findings, it is helpful to recall the mechanism which underlies our predictions. It argues that if advisers base what they consider impartial advice on their own preferences, then they have to act according to their biased, previous advice in order to not signal the fact that they were

corrupted. Therefore, the root cause for the persistent effect on the adviser's own choice is that the bonus led advisers to recommend option A in the first recommendation. This initial bias then affects the subsequent own choice in O. To investigate the mediating effect of the first recommendation we also estimate the above regression model when an indicator for whether the first recommendation was for option A (i.e. the dependent variable from model 1) is included as an additional independent variable. If the bonus worked via the initial recommendation, its effect should be captured by the coefficient on this additional regressor. Column 4 in table 3 shows that exactly this happens: The previously positive and statistically significant coefficient on $Bonus_i$ becomes essentially zero and insignificant while the coefficient on $r_{1,i} = A$ takes up all the explanatory power. The point estimate implies that having initially recommended option A significantly increases the probability of later choosing it for oneself by 38.7 percentage points.

We can also quantify more exactly the effect which the bonus had on advisers' own choices when their previous advice was biased by it. To form an estimate of this conditional effect, we divide the unconditional effect of the bonus on own choices by its effect on the initial recommendations. This assumes that this channel is the only way how the bonus affected subsequent own choices. From the results above, we get that the increase in O equals 17.3 percentage points while the increase in R1 is 50.3 percentage points, based on unconditional mean differences. We then get from these numbers that 34.4% ($\triangleq$ 0.173/0.503) of the advisers whose initial advice was shifted towards option A by the bonus also adjusted their own choices accordingly.

Note that the above estimate is the Wald/grouping-estimator one would also obtain in the second stage of a 2SLS-estimation without further controls. In this regression, random assignment to the BONUS-treatment would be first used to predict recommendations for option A in R1 and then, based on these predictions, the effect of the initial recommendation on own choices would be estimated in the second step. In column 5 of table 3 we present the second-stage results when additional controls are added (the corresponding first stage results are thus reported in column 3 in table 2). When we add these controls, the mediating effect of the initial recommendation, modeled explicitly through the initial presence of the bonus, increases to a 45.3 percentage-point-shift in the probability of later choosing the risky option for oneself. Note that capturing the mediating effect of the initial recommendation also increases the explanatory power of the regression model. This is documented by the increase in the adjusted R^2 in column 4 and 5, relative to columns 1 through 3 where the mediating effect is not explicitly modeled.

Results for R2: For the second recommendation, the decision situation for advisers in NO BONUS is the same as for their first recommendation. Accordingly, we expect a similar pattern of recommendations in this treatment. The left panel of figure 3 shows the recommendation frequencies for each

option. In fact, only a small fraction of advisers in NO BONUS recommended option A – exactly

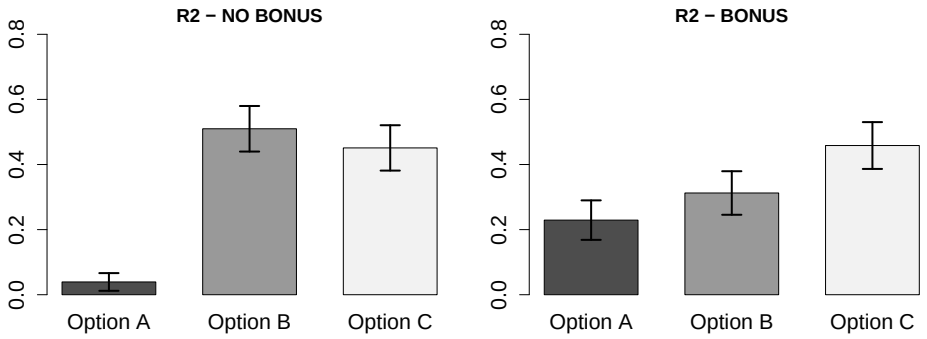


Figure 3: Frequency for each option being recommended in R2, together with standard errors.

the 3.9% who also recommended this option previously in R1 (see also table 6a below). The right panel of this figure, which displays the distribution of recommendations in BONUS, is very different. Although there is no bonus in R2 anymore, the rate of recommendations for option A is almost six times as high as in NO BONUS: 22.9% of those advisers who had previously been exposed to the bonus recommended option A, a significant increase by 19.0 percentage points (Fisher exact test: $p = 0.007$). As before, we perform a regression analysis by estimating model (1), now with a dummy which indicates whether option A is recommended in the second recommendation as the dependent variable. Columns 1 through 3 in table 4 present the results and show that the estimate for the effect of having been offered the bonus remains of roughly the same size and is still significant when controls are added. These results therefore support Prediction 3.

	(1)	(2)	(3)	(4)	(5)
$Bonus$	0.211**	0.192***	0.213**	0.025	
	(0.092)	(0.067)	(0.087)	(0.091)	
$r_1 = A$				0.390***	
				(0.123)	
$\widehat{r_1 = A}$ (via $Bonus$)					0.442***
					(0.142)
Personal Controls	yes	no	yes	yes	yes
Session Controls	no	yes	yes	yes	yes
Observations	99	99	99	99	99
Adjusted R^2	0.038	0.061	0.064	0.241	0.248

Table 4: OLS/2SLS estimates of the probability to recommend option A in R2.
*Robust standard errors in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.*
Personal controls: age, gender, monthly budget, subject's region of origin, degree and field of studies.

As for the own choice O, we also checked for the mediating effect the bonus had on the second recommendation through the first recommendation. If one includes an indicator for $r_{1,i} = A$ as an additional independent variable, its effect is highly significant while the coefficient of the bonus-dummy becomes almost zero and insignificant (see column 4 in table 4). We also calculated the associated share of advisers who re-recommended option A because they have initially recommended

it for the bonus. Based on unconditional mean differences we get for the resulting Wald/group-estimator, via the same procedure as in O, that 37.8% ($\triangleq$ 0.190/0.503) re-recommended option A because the bonus affected their first recommendation. In column 5 of table 4 we report the corresponding estimate when we control for observables, i.e. the second stage 2SLS-coefficient on having previously recommended option A when this independent variable is predicted by assignment to the BONUS-treatment in the first-stage. Again, adding these controls increases this implied percentage of advisers to 44.2%. We thus find that the share of advisers who re-recommended option A because of their initial bias is in the same range as the corresponding share who chose so for themselves, between 35% and 45%.

Further results: There are some additional findings which support our theory and its underlying assumptions. Given our previous results, we expect consistency between advisers' own choices and their first recommendation when there is no conflict of interest. Our results support this. Table 5a shows the frequency of advisers' own choices in O, conditional on their initial recommendations in R1 for NO BONUS. Only the off-diagonal entries are not predicted. They amount to a total of

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

a) NO BONUS

R1 \ O	A	B	C
A	22.9%	8.3%	22.9%
B	0.0%	6.3%	0.0%
C	4.2%	4.2%	31.2%

b) BONUS

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

17.7% of the observations in this treatment; 82.3% of our observations in NO BONUS are therefore in line with the predicted consistency. For BONUS, our theory predicts that some of those who have previously recommended option A stick to it in order to avoid a negative self-image. Other advisers who have recommended it but who did not have sufficiently strong image concerns chose 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, 8.4% of our observations, unexplained.

We find similar results with regards to the consistency between advisers' first and second recommendations. The two panels of table 6 show the conditional frequencies across our experimental conditions. In NO BONUS, we observe that 17.7% of the second recommendations are inconsistent with the first recommendation, i.e. are outside table 6a's diagonal. All of them are, however, switches from having initially recommended option C and then option B, but not switches to or from the risky option A. With regards to variations in the BONUS treatment, the results are even stronger. In total, 87.5% of its observations fall into an explainable pattern, thus are either on the

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

a) NO BONUS

R1 \ R2	A	B	C
A	18.7%	16.7%	18.8%
B	0.0%	6.2%	0.0%
C	4.2%	8.3%	27.1%

b) BONUS

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

diagonal or the first row of table 6b. Overall, we find that, in terms of consistency, more than four out of five observations follow a pattern predicted by our theory.

Further evidence comes from our exit questionnaire. Besides questions asking for the subjects' personal characteristics, it also 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?". This question was not incentivized, but answers to it have previously been shown to correlate with peoples' incentivized choices under risk (see Dohmen et al., 2011). While in NO BONUS, the average response was 5.0 points, this increased by almost one point or alternatively, 39.8% of the measure's standard deviation, to 5.9 points in the BONUS-treatment. In a non-parametric test, this increase is marginally statistically significant (Wilcoxon ranksum-test: $p = 0.059$).[15] These results become even stronger, both numerically and statistically, if we control for individual characteristics. Columns 1 through 3 of table 7 present the results from estimating model (1) when the dependent variable is the self-assessed risk-measure and controls are successively added.

	(1)	(2)	(3)	(4)	(5)
$Bonus$	1.049**	0.891*	1.074**	0.236	
	(0.456)	(0.449)	(0.468)	(0.479)	
$r_1 = A$				1.739***	
				(0.472)	
$\widehat{r_1 = A}$ (via $Bonus$)					2.230***
					(0.806)
Personal Controls	yes	no	yes	yes	yes
Session Controls	no	yes	yes	yes	yes
Observations	99	99	99	99	99
Adjusted R^2	0.232	0.048	0.207	0.278	0.282

Table 7: OLS/2SLS estimates of the self-assessed preference for risk (Likert scale, 0 - 10).
*Robust standard errors in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$*
Personal controls: age, gender, monthly budget, subject's region of origin, degree and field of studies.

[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 data 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 results of the regression analysis presented in table 7 remains valid with higher point estimates and slightly higher standard errors in a way such that p-values decrease. Note that our primary data on the recommendations in R1/R2 and on own choices in O were not affected by this data glitch since advisers wrote them on paper.

This increase in an adviser's self-stated risk measure is consistent with our theory and our previous findings for advisers' own choices. Advisers who have previously given in to the bonus can signal that this advice was appropriate, from their point of view, when they consider themselves as more risk-seeking. Preceding as before to check for such a mediating effect, we find that if a dummy indicating whether option A was recommended in R1 is added as an independent variable, the previously positive and significant coefficient for $Bonus_i$ vanishes. The point estimate implies that when option A was initially recommended, the self-stated risk preference significantly increases by about 1.7 points (column 4 in table 7). Under the assumption that only the bonus' effect on the initial recommendation caused this shift, one can again compute the bonus' effect on those whose advice it biased. The Wald/grouping estimator, based on the unconditional mean shifts, implies that the self-stated risk preference for such advisers increases by 1.8 ($\triangleq$ 0.090/0.503) points. As before, the corresponding 2SLS-estimate of the bonus-mediated effect increases if one adds additional controls. Its estimate corresponds to 2.2-point shift in the self-stated preference for risk for those whose initial advice was biased towards option A by the bonus (column 5 in table 7). We therefore observe that the bonus' influence on initial recommendations for the risky choice did not only affect advisers' further recommendations and choices from the same set of lotteries but also their answers to a more general question with regard to risk.

6 Discussion & Conclusion

In this paper, we present evidence that incentives to bias advice have a lasting and causal effect on both, advisers' future recommendations for risky decisions and their own choices. We estimate that 35% to 45% of the advisers whose initial advice was biased by a bonus to recommend a very risky lottery recommended it again in a second advice-giving situation. They did so, even though they did not receive a bonus for the second recommendation. This is consistent with our proposed mechanism which captures the insight that changing advice after a conflict of interest disappeared signals one's prior corruptibility. This mechanism assumes that issuing biased advice is costly. Only when these costs are low enough, relative to image costs, advisers stick to their initial, biased recommendation. In line with this, these estimates imply that two-thirds to one-half of the advisers whose initial advice was biased by the bonus changed their recommendation after the bonus was removed. An alternative explanation in which advisers are clueless about which option to take and then just take the bonus as a cue for what is the "best" option makes different predictions. In this case all advisers whose initial advice was affected by the bonus should stick to their cue-induced initial recommendation. The *partial* consistency we find speaks against such an anchoring effect.

We also find similar effects with regards to advisers who chose the risky lottery for themselves because the bonus induced them to initially recommend it. This allows us to differentiate how advisers determine appropriate advice and what signals unbiased advice. If advisers had formed a motivated belief about their clients' preferences independent of their own, this would not have required them to choose consistently for themselves in order to avoid signaling their corruptibility. If impartial advice however relates to what one would choose for oneself, it is necessary to choose according to the previous advice for oneself in order to avoid such a signal. The results suggest that the latter reasoning is relevant. The fact that consistency for option A in own choices is also only partial, again speaks against an anchoring effect. These findings rather indicate a trade-off – this time between image costs and the costs of not choosing one's actually preferred choice.

Further evidence in this direction comes from the answers to the question on general risk-taking. We present evidence that having recommended the risky option for a bonus led advisers to state a higher general preference for risk. This shows that they did not only act consistently in a mechanical way when there is a fixed set of possible lotteries. Rather, it shows that the signaling implications of their initial, biased recommendations apply to wider, but related, choices.

These results can be explained by the mechanism we propose and in which image concerns are the crucial ingredient. Such concerns could be driven by advisers trying to maintain a positive self-image as in dual-self models. To capture the notion that one constantly learns through one's actions about the own underlying motives, these models feature a learning self who infers as if it were an outside observer. Social image concerns with regard to an actual outside observer would therefore lead to the same effects as long as this observer sees the sequence of an adviser's choices and recommendations. In our experiment, social image concerns were limited by the fact that advisers wrote down their recommendations and choices in private and put them in envelopes. Thus, even the experimenter could not be expected to see the sequence of advisers' actions and choices during the experiment (except by the low probability-event of repeatedly sampling a given adviser's envelopes for implementation). While we therefore interpret our findings to be caused mainly by self-image concerns, social image may also matter in many important situations. For example, supervisors or regulators regularly observe the sequence of advisers' recommendations before and after conflicts of interest were present and social image concerns towards them can even amplify the effects presented here.

The findings we present have several immediate implications. First, we present evidence that biases in advice-giving loom longer than the conflict of interest which caused them. Recent policies which ban the causes of conflicts of interest are certainly a right step towards eventually achieving impartial advice. However, our results show that they should not be taken as a guarantee that advice

becomes immediately impartial. Second, our proposed mechanism and the results in line with it show that a desire to be perceived as behaving impartially can have the consequence of inducing exactly the opposite behavior. This can be especially relevant in situations where experienced advisers have long been exposed to such incentives. Changes in their recommendations can then constitute a considerable threat to their professional identity by indicating a previous bias. Also, if such a reasoning happens sub-consciously then advisers believe that they act in a morally intact manner even though they are actually biased. Investigating whether advisers are aware of the initial bias and its persistence might therefore be important to judge its effect and whether they can eventually be held accountable. Third, we find a persistent effect on advisers' second recommendations after incentives to bias them were removed, even though they had to choose for themselves before. This observation speaks against the potential of requiring advisers to choose for themselves in having a "cleansing effect" on subsequent advice. Finally, it is important to note that while our findings are on advice for risky choice, they are not necessarily bound to this specific domain. The crucial feature is that there is no clear-cut right or wrong so that one can reasonably maintain the image that advice was genuine and unbiased, even though it was not. Similar effects could therefore also be found for advice on moral, legal or other complex decisions. Investigating these domains would provide other avenues for further research.

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

In the following, we set up a formal model which demonstrates how advisers can be affected by conflicts of interest, even after they have been removed. The key assumptions underlying it reflect those described in Section 2. In the model, we establish Corollary 1 through 3 which are analogous to the respective predictions in the main text.

A1. Model setup

We consider an adviser who advises a client on which option out of a discrete, finite set of choices $\mathcal{C}$ to take. The adviser may also have to choose for himself from this set. He gets a bonus payment $b \geq 0$ if he recommends a choice from the set $\mathcal{B} \subset \mathcal{C}$. Hence, when $b > 0$, the adviser is subject to a bias towards recommending a choice from $\mathcal{B}$.[16] Three factors influence an adviser's actions: 1) his own (expected) pecuniary payoff, 2) costs of giving inappropriate advice, and 3) image concerns of being perceived of having given biased advice. We will explain them in detail below:

Adviser's payoff and personal preferences: We assume that pecuniary payoffs map into the adviser's utility via the strictly increasing vNM-utility function $u : \mathbb{R} \to \mathbb{R}$ with $u(0) = 0$. We also assume that all choices in $\mathcal{C}$ can be represented by a pecuniary payoff. This payoff is captured by the function $v : \mathcal{C} \to \mathbb{R}$, for example via the corresponding certainty equivalent when elements of $\mathcal{C}$ are lotteries. We omit the v-function if it is the identity function.

The own choice o which an adviser optimally chooses for himself from a, possibly restricted, subset $\mathcal{X} \subseteq \mathcal{C}$ is denoted by $o^*_{\mathcal{X}} \equiv \max_{o \in \mathcal{X}}\{v(o)\}$. To save on notation we assume w.l.o.g. that $o^*_{\mathcal{X}}$ is a singleton for each $\mathcal{X} \subseteq \mathcal{C}$. The subscript is omitted when the choice-set is unrestricted, i.e. $o^* = o^*_{\mathcal{C}}$. The share of advisers for whom $o^* \in \mathcal{X} \subseteq \mathcal{C}$, i.e. whose unconstrained optimum lies in $\mathcal{X}$ (for such advisers $o^* = o^*_{\mathcal{X}}$), will be denoted with $\alpha_{\mathcal{X}}$.

Costs of giving inappropriate advice: Each adviser has a single "advisable choice" of which he thinks that it should be recommended to the client. This ideal recommendation is denoted by $r^* \in \mathcal{C}$. We denote with $\beta_{\mathcal{X}}$ the share of advisers for whom $r^* \in \mathcal{X} \subseteq \mathcal{C}$, i.e. those who think that the option they consider advisable is in $\mathcal{X}$. In the following, we will consider two prominent possibilities of how r^* is determined:

- *Projected advisable recommendations:* As mentioned in the main text, there is ample evidence that people project their own preferences onto others, e.g. the client. Equivalently, they might follow a general rule which stipulates, for themselves and others equally, what ought to be chosen. This would mean that $r^* = o^*$ holds and therefore $\alpha_{\mathcal{X}} = \beta_{\mathcal{X}}$ for each $\mathcal{X} \subseteq \mathcal{C}$.

[16] Note that this is equivalent to a punishment $p = -b$ he has to pay if he does not recommend an option from $\mathcal{B}$.

- *Independent advisable recommendations:* Alternatively, advisers may base advisable choice on criteria which are unrelated to their own preferences. For example, they can hold an independent belief about the client's preferences which then stipulates which choice would suit her best. Importantly, such a belief can be motivated and instrumental in helping the adviser to recommend a choice he would not prefer for himself. In such a setting, o^* and r^* and their respective distributions are independent.

Giving inappropriate advice creates costs for the adviser. In the context of our experiment these costs are psychological but they could also be expected legal costs or both. They are captured by the dis-utility $\kappa \geq 0$ which an adviser experiences if he recommends an option $r \in \mathcal{C}$ when $r \neq r^*$.

Image costs of being perceived as biased: In addition to the immediate costs of not recommending an advisable choice, we also allow for costs of being *perceived* ex-post of having acted in such a manner. More precisely, we assume that the adviser experiences dis-utility $\lambda \geq 0$ to the degree that he (or someone else who observes his actions) learns that previous advice was biased, i.e. that a previous recommendation r did not correspond to the adviser's advisable action r^*. This "degree" which weighs these costs corresponds to the posterior probability that, given an adviser's prior and current actions, previous advice was biased. The observer who makes such an inference observes the adviser's actions but not his advisable action r^*. Given our setup and findings, we interpret such image concerns as self-image concerns. This corresponds to a dual-self model, similar to Bodner and Prelec (2003) or Bénabou and Tirole (2011): One self is a standard economic agent who trades off the benefits and costs of any action and knows, e.g. via a gut-feeling, whether the adviser gave inappropriate advice or not. The other self does not know this and is modeled as an outside observer who sees an adviser's actions. Thus, *social* image concerns regarding an actual outside observer follow the same model.

Payoff and utility function: Let $\pi(a)$ denote the corresponding pecuniary payoff which an adviser gets from action $a \in \mathcal{C}$ which is either a recommendation or a choice for himself. Specifically,

$$\pi(a) = \begin{cases} v(a) & \text{if } a = o \text{ is the adviser's own choice for himself;} \\ b \cdot \mathbb{1}[a \in \mathcal{B}] & \text{if } a = r \text{ is the adviser's recommendation to a client.} \end{cases}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function which takes a value of one if the statement in the bracket is true. Given an adviser's history h of previous own choices and recommendations plus the choice r^* he considers advisable, his overall utility can then be written as follows:

$$\begin{aligned} U(a \mid h, r^*) = u(\pi(a)) &- \kappa \cdot \mathbb{1}[a \neq r^* \text{ and } a \text{ is a recommendation}] \\ &- \lambda \cdot \Pr[\text{previous advice was biased} \mid a, h] \end{aligned} \tag{2}$$

In the above, the first term denotes the adviser's utility from pecuniary payoffs. The second term denotes the costs of not giving an appropriate recommendation. The third term is the expected image cost of being perceived as biased. Note that when $\Pr[$*previous advice was biased* $\mid a, h]$ varies with changes in a ("a has diagnostic value"), choosing an action a which maximizes (2) corresponds to playing a signaling game. To solve such a game, we use Perfect Bayesian Equilibrium as a solution concept: Given a prior history h, the adviser chooses a such that $U(a \mid h, r^*)$ is maximized given that $\Pr[$*previous advice was biased* $\mid a, h]$ is updated via Bayes' rule under knowledge of the adviser's strategy. We focus on pure strategies. As a tie-breaking rule we make the (natural) assumption that if an adviser is indifferent between multiple choices for himself including o^*, his preferred choice, he chooses o^*. Similarly, if he is indifferent between recommending different choices of which one is r^*, the choice he considers advisable, he recommends r^*.

Heterogeneity and information structure: Advisers' moral and image costs are heterogeneous. We denote the corresponding joint distribution via its c.d.f. $J(x, y) = \Pr[\kappa \leq x, \lambda \leq y]$. To save considerably on notation, we assume that κ and λ are independent of o^* and r^*.[17] Note that this does not prevent λ and κ to be correlated. We can then state the following:

Lemma 1. *Suppose the joint distribution of (κ, λ) is absolutely continuous and the associated p.d.f. has full support over $\mathbb{R}_0^+ \times \mathbb{R}_0^+$. Then, the following holds:*

a) The marginal c.d.f.s $K(x) = \Pr[\kappa \leq x]$ and $\Lambda(y) = \Pr[\lambda \leq y]$ are strictly increasing for every $x, y \geq 0$.

b) The conditional marginal c.d.f. $\Lambda(y \mid x) = \Pr[\lambda \leq y | \kappa \leq x]$ is strictly increasing in y for every $y \geq 0$ and for any $x > 0$.

c) The conditional c.d.f. for the distribution of the ratio $(\kappa/\lambda \mid \lambda > 0)$, given by $R(z \mid x, y) = \Pr[\kappa/\lambda \leq z \mid \kappa \leq x, \lambda \geq y]$, exists and is non-decreasing in z for every $x, y > 0$.

Proof: see Appendix B.

[17] With such correlation all our results would remain valid if the joint distributions of (κ, λ), conditional on preferred own choice o^* and advisable choice r^* are increasing. For example, full support for $J_c(x, y) \equiv \Pr[\kappa \leq x, \lambda \leq y \mid o^* = c]$ and $\tilde{J}_c(x, y) \equiv \Pr[\kappa \leq x, \lambda \leq y \mid r^* = c]$ for all $c \in \mathcal{C}$ would be such a sufficient (but not necessary) condition.

In the following, we assume that the above assumptions and therefore Lemma 1 hold.[18] We also assume that the joint distribution J, together with the families of distributions $\{\alpha_{\mathcal{X}}\}_{\mathcal{X}\subseteq\mathcal{C}}$ and $\{\beta_{\mathcal{X}}\}_{\mathcal{X}\subseteq\mathcal{C}}$ which describe the distribution of advisers' preferences and what they consider advisable choices, are common knowledge. To make things interesting, we assume that some but not all advisers consider an option advisable which would not earn them the bonus, i.e. $\beta_{\mathcal{C}\setminus\mathcal{B}} \in (0,1)$. While these distributions are common knowledge and the adviser, when he chooses a to maximize (2), knows (κ, λ, r^*), the observer or the observing self does not know these parameters.

A2. Analysis

We will now analyse how a one-off incentive can lead to a persistent bias in advice-giving. For this, we consider a situation in which an adviser first has to issue a recommendation $r_1 \in \mathcal{C}$ for which he can earn a bonus b and then a second recommendation $r_2 \in \mathcal{C}$ to another client for which no bonus can be earned.

Our experiment resembles this setting in the BONUS-treatment where $b = 3\,GBP$. It also includes a counter-factual where no incentive to bias advice is ever present, the NO BONUS-treatment with $b = 0\,GBP$. In the experiment, $\mathcal{C} = \{A, B, C\}$ and $\mathcal{B} = \{A\}$ hold. In addition to a second recommendation stage, our experiment also features a stage where, after having made the first recommendation but before the second, the adviser has to make an own choice $o \in \mathcal{C}$ for himself. For this own choice, no bonus can be earned either. This allows us to separate whether advisers form motivated beliefs or whether they tie advice to own preferences which prevents such self-serving beliefs. However, as will become clear, the main result regarding the persistent bias in advice-giving is independent of whether there is an own choice or not.

Advisers' behavior is analyzed step by step, in the order as subjects acted in the experiment: We start with the first recommendation (R1), then treat the own choice (O), and finally cover the second recommendation (R2). In each step we contrast behavior when there was an initial conflict of interest (BONUS) with behavior when there was no such conflict (NO BONUS).

First recommendation R1

R1 – NO BONUS: Here, the adviser's action a is a recommendation denoted by $a = r_1$. There is no prior advice and therefore, image concerns do not matter. Using that $\pi(r_1) = 0$ because $b = 0$, (2) becomes $U(r_1 \mid r^*) = -\kappa \cdot \mathbb{1}[r_1 \neq r^*]$. Accordingly, advisers recommend $r_1 = r^*$ and the share of advisers who recommend an option from $\mathcal{B}$ is given by $\beta_{\mathcal{B}}$.

[18] Precluding the possibility of mass at $\kappa = 0$ and/or $\lambda = 0$ is done to save on additional notation. The model can be extended to accommodate this possibility without any of the results being changed.

R1 – BONUS: Recommending an option from $\mathcal{B}$ now yields the bonus, captured by $\pi(r_1) = b \cdot \mathbb{1}[r_1 \in \mathcal{B}]$ with $b > 0$. There is no previous advice, so that image concerns do not matter, thus (2) becomes $U(r_1 \mid r^*) = u(b \cdot \mathbb{1}[r_1 \in \mathcal{B}]) - \kappa \cdot \mathbb{1}[r_1 \neq r^*]$. For the share $\beta_{\mathcal{B}}$ of advisers who have $r^* \in \mathcal{B}$, recommending $r_1 = r^*$ is then clearly optimal – they get rewarded for what they would have recommended anyway. However, for a share $1 - \beta_{\mathcal{B}}$ of advisers, $r^* \notin \mathcal{B}$, holds. They face a trade-off between recommending an option from $\mathcal{B}$ even though they do not consider it advisable and being impartial. Recommending an option from $\mathcal{B}$ yields them pecuniary utility $u(b)$ but causes costs κ of giving inappropriate advice. Being impartial by recommending $r_1 = r^* \notin \mathcal{B}$ does not create such costs but no bonus is earned either. Accordingly, those with costs κ lower than $u(b)$ give biased advice; their population share is given by $(1 - \beta_{\mathcal{B}}) \cdot K(u(b)) > 0$. It follows that advisers' behavior in the BONUS-treatment corresponds to three different behavioral types, denoted by $\theta \in \{1, 2, 3\}$ which are determined by their values for (κ, λ) and r^*:

- $\theta = 1$ – incorruptible advisers who recommend an option which earns them a bonus because they truly think that it is advisable for the client ($r_1 = r^* \in \mathcal{B}$). Their population share is $\phi_1 \equiv \beta_{\mathcal{B}}$.
- $\theta = 2$ – incorruptible advisers who recommend an option which does not earn them a bonus because they think that it is advisable ($r_1 = r^* \notin \mathcal{B}$) and who are not corrupted by the bonus because their κ is sufficiently high. Their population share is $\phi_2 \equiv (1 - \beta_{\mathcal{B}}) \cdot (1 - K(u(b))) > 0$.
- $\theta = 3$ – corruptible advisers who recommend an option which earns them a bonus though they do not think that it is advisable ($r_1 \in \mathcal{B}$ but $r_1^* \notin \mathcal{B}$). They are corrupted by the bonus because their κ is low enough. Their population share is $\phi_3 \equiv (1 - \beta_{\mathcal{B}}) \cdot K(u(b)) > 0$.

Note that by letting $b = 0$, the above also applies to the NO BONUS-treatment. In this case, only share ϕ_1 recommends an option from $\mathcal{B}$ as there are no type-3-advisers. Also note that from the above, $\Pr[\textit{previous advice was biased} \mid a, h] = \Pr[\theta = 3 \mid a, h]$ holds. We then immediately get the following corollary which resembles Prediction 1 in the main text:

Corollary 1. *The share of advisers who recommend a choice from $\mathcal{B}$ in BONUS is given by $\phi_1 + \phi_3$ and is larger than the share ϕ_1 of advisers who recommend such a choice in NO BONUS.*

Own choice O

O – NO BONUS: Now, the action a in (2) is the adviser's own choice for himself and denoted by $a = o$. Accordingly, its corresponding pecuniary value can be expressed by $\pi(o) = v(o)$. As this is not an advice to a client, no cost κ of giving inappropriate advice matters. Without a bonus, only type-1 and type-2-advisers exist so that there are no concerns of being perceived as biased.

Therefore, (2) becomes $U(o \mid r_1, r^*) = u(v(o))$ which is maximized by an adviser's preferred own choice o^*. The share of advisers choosing an option from $\mathcal{B}$ is thus given by $\alpha_{\mathcal{B}}$.

O – BONUS: As there were corruptible advisers in the previous recommendation R1 (the type-3-advisers) image costs of being perceived as them matter. Given the prior history $h = r_1$ and the current choice for oneself $a = o$ the adviser's objective function (2) then becomes $U(o \mid r_1, o, r^*) = u(v(o)) - \lambda \cdot \Pr[\theta = 3 \mid r_1, o]$. It therefore matters whether o has diagnostic value:

- *Projected advisable recommendations:* This means that $r^* = o^*$. An intuitive implication is then that, were it not for the bonus, advisers should choose what they have recommended. Type-3-advisers who have previously recommended $r_1 \neq o^* = r^*$ would be put on the spot: By choosing $o = o^* \neq r_1$ they would reveal themselves as type-3-advisers with $r_1 \neq r^*$ because type-1 and type-2-advisers choose $o = o^* = r_1 = r^*$. Type-3-advisers would then suffer full dis-utility λ for the benefit of choosing their own preferred choice. Alternatively, type-3-advisers could pool with type-1-advisers by choosing $o = r_1 \in \mathcal{B}$. By this, they would lower the weight on the image costs of being perceived as corruptible but incur costs of not choosing what they actually prefer as for them, $o^* \notin \mathcal{B}$ holds. The following proposition shows that such behavior is indeed the unique equilibrium in this situation and that some, but not all, type-3-advisers choose a non-preferred choice for themselves to pool with type-1-advisers:

  Proposition 1. *When $b > 0$ and $o^* = r^*$, there is a unique equilibrium in which all advisers of type $\theta \in \{1, 2\}$ and a share $\pi_o^* \in (0, 1)$ of advisers of type $\theta = 3$ choose $o = r_1$. The remaining share of type-3-advisers chooses $o \neq r_1$.*
  Proof: see Appendix B.

- *Independent advisable recommendations:* Having a (possibly self-serving) belief about what constitutes the advisable choice such that r^* and o^* are independent prevents the above-described pressure on type-3-advisers. Such a belief prevents any inference via o on whether $r_1 = r^*$ holds, i.e. $\Pr[\theta = 3 \mid r_1, o]$ is invariant to o. The choice $o \in \mathcal{C}$ which maximizes $U(o \mid r_1, o, r^*) = u(v(o)) - \lambda \cdot \Pr[\theta = 3 \mid r_1, o]$ is thus the one which maximizes its first element, given by o^*.

The following corollary then follows directly from the above and describes the situation for own choice O across the two environments with and without a bonus:

Corollary 2. *When own choices and advisable choices are identical (projected advisable choice), the share of advisers choosing $o \in \mathcal{B}$ in BONUS is given by $\phi_1 + \pi_o^* \phi_3$ with $\pi_o^* \in (0, 1)$ and is larger than ϕ_1, the share of advisers who recommend such a choice in NO BONUS.*

When own choices and advisable choices are independent (independent advisable choice), the share of advisers choosing $o \in \mathcal{B}$ is given by ϕ_1 in both, BONUS and NO BONUS.

Second recommendation R2

R2 – NO BONUS: Here, a is another recommendation, denoted by $a = r_2$. As with the first recommendation in NO BONUS, there is no payment involved, thus $\pi(r_2) = 0$. Also, there are no type-3-advisers in this condition so that image concerns do not matter. Since r_2 is a recommendation, costs of giving inappropriate advice remain and (2) becomes $U(r_2 \mid r_1, r_2, r^*) = -\kappa \cdot \mathbb{1}[r_2 \neq r^*]$. Recommending $r_2 = r^*$ maximizes this and share ϕ_1 of advisers recommends an option from $\mathcal{B}$.

R2 – BONUS: In this condition, the initial bonus has been removed so that $\pi(r_2) = 0$ holds, as in NO BONUS. However, as there was a bonus in the first recommendation, there is a positive mass of type-3-advisers and image concerns matter, in addition to the costs of giving inappropriate advice. The advisers' utility (2) then becomes $U(r_2 \mid r_1, r_2, r^*) = -\kappa \cdot \mathbb{1}[r_2 \neq r^*] - \lambda \cdot \Pr[\theta = 3 \mid r_2, r_1, o]$.

Similar to own choices under projected advisable recommendations, this puts type-3-advisers on the spot again: If advice in the initial recommendation was unbiased, this advice should be issued again as the presence of a bonus should not have affected the initial recommendation r_1. Accordingly, type-1 and type-2-advisers should recommend $r_2 = r_1$. Type-3-advisers who have not revealed themselves as such yet could then re-recommend $r_2 = r_1 \in \mathcal{B}$ in order to pool with type-1-advisers which discounts their image costs λ. However, since for them $r^* \notin \mathcal{B}$ holds, they would then suffer costs κ of issuing biased advice (again). If they want to prevent these costs and recommend $r_2 = r^*$, this means that $r_2 \neq r_1$ and they would therefore reveal themselves as type-3-advisers and suffer full image costs λ. They do so if λ is small, relative to the costs κ of re-issuing biased advice.

The above reasoning applies to type-3-advisers who have not yet revealed themselves. This is always the case when advisable recommendations are independent. In contrast, when advisable recommendations are projected, some type-3-advisers have already revealed themselves by choosing $o \neq r_1$. For them, there is no point of trying to pool with type-1s. However, as Proposition 1 shows, there is then always a non-zero share π^*_o of type-3 advisers who have not yet revealed themselves and to whom the above reasoning applies. Proposition 2 summarizes this and proves that the partial pooling described above is the unique equilibrium outcome which leads to the following corollary.

Proposition 2. *When $b > 0$, there is a unique equilibrium in which all advisers of type $\theta \in \{1, 2\}$ and share $\psi \cdot \pi^*_{r_2}$ of advisers of type $\theta = 3$ choose $r_2 = r_1$, the other advisers with $\theta = 3$ recommend $r_2 \neq r_1$. For this, it always holds that $\pi^*_{r_2} \in (0, 1]$. If $o^* = r^*$ (projected advisable recommendations), then $\psi = \pi^*_o$. If o^* and r^* are independent (independent advisable recommendations), then $\psi = 1$.*
Proof: see Appendix B.

Corollary 3. *The share of advisers re-recommending a choice from $\mathcal{B}$ in BONUS is given by $\phi_1 + \psi\pi^*_{r_2}\phi_3$ and is larger than ϕ_1, the share of advisers who recommend such a choice in NO BONUS.*

A3. Discussion of the model

The above analysis shows that, through image concerns of being perceived as biased, a one-off bonus can lead advisers to repeat biased advice even though there is no bonus anymore. It can even lead them to choose for themselves in a way which, absent such concerns, would be sub-optimal. For this, an adviser has to be initially influenced by a previous bonus, thus he has to have small enough costs κ of giving inappropriate advice. The persistent effect then occurs when, in addition to sufficiently low values of κ, image costs, as measured by λ, are high enough.

The grey rectangle in figure 4a depicts the relevant parameter constellations such that own choices are persistently affected by the bonus: Low enough costs to bias advice (κ below the vertical line) and high enough image concerns (λ above the horizontal line). Figure 4b shows the parameter-values which lead to persistent bias in advice-giving. For this, advisers have to have sufficiently high image concerns such that they rather re-recommend an in-advisable choice than to change their advice and thereby reveal themselves as corruptible. Accordingly, image costs λ have to be high relative to the costs κ of issuing inappropriate advice, i.e. above the bold diagonal line. When own choices are diagnostic (projected advisable recommendations), this is only relevant to those in the left grey rectangle who have not yet revealed themselves. In this case, only those who have these parameters plus a sufficiently high ratio λ/κ will re-issue biased advice. Therefore, the relevant

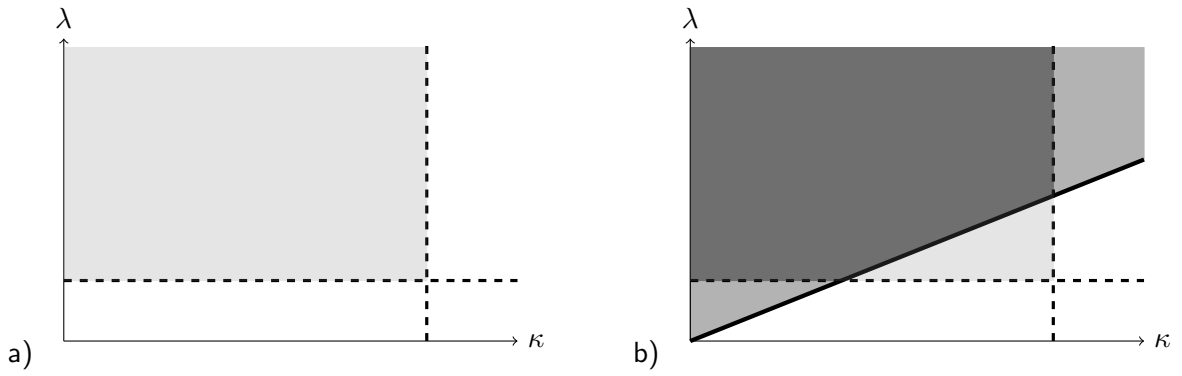


Figure 4: (κ, λ)-values which imply persistent bias in own choice (light grey rectangle) and recommendations (dark grey pentagon when own choices have diagnostic value, grey area above diagonal when not).

parameters are within the dark grey pentagon. If own choices have no diagnostic value (independent advisable recommendations) the parameter restrictions displayed in the left panel can be ignored and it is only the right panel's diagonal which sets the threshold for repeated biased advice. Every advisor with values of (κ, λ) above the diagonal then re-issues biased advice. Note that the latter reasoning would also apply if there were only repeated advice-giving and no own choice.

In the model, image costs are defined rather broadly. They are measured by a variable λ, scaled by the inferred probability that initial advice was biased as evaluated when one looks on an adviser's current and prior actions. We interpret this as self-image concerns within a dual-self model because

our experimental setup tried to minimize social signaling concerns. However, for the main effect we present and the above theoretical mechanism behind it, the source of image concerns is irrelevant so that social image concerns have the same effect. The only relevant feature is that the party whose image of the adviser is important to him can observe the history of his actions.

It also deserves discussion that we model image concerns as being perceived as corrupted in *past* advice. Accordingly, such concerns do not play a role in the initial recommendation when there is no past behavior. We believe that this is sensible in the present context. The reason is that the possibility to reveal oneself as an adviser who has given biased advice only occurs when the bonus is removed and one acts inconsistently afterwards. However, when advice was first given in R1, advisers in our experiment did not know that there would be a possibility to act inconsistently. Similarly, in the real-world settings we have in mind, advisers give biased advice until, through some unexpected exogenous events such as a scandal and resulting new regulations, this environment is reshaped by removal of the conflict. In addition, the costs κ already capture the immediate (expected) costs of giving biased advice. The image costs λ therefore capture the additional costs of being reminded ex-post, through one's subsequent actions, that previous advice was biased.

In principle, image concerns with respect to the current actions could be introduced. However, this would lead to a less tractable model and, arguably, also less realism. To see this, suppose image concerns were with respect to past and current behavior. The three behavioral types presented in the above analysis would persist. In addition, there could be a fourth type of advisers. This type would form a subgroup of the type-1s who think that an option from $\mathcal{B}$ is advisable. However, due to immediate image concerns, these type-1s recommend an option from $\mathcal{C} \setminus \mathcal{B}$ to avoid the perception that they are biased. In later stages, some of these advisers would then switch back to choose or recommend their preferred choice from $\mathcal{B}$ while others, similar to type-3-advisers, would stick to their biased recommendation from $\mathcal{C} \setminus \mathcal{B}$, just to signal that their initial advice was not biased (although it is actually biased, but *away* from the bonus). Thus, this bias is not directly caused by the bonus but indirectly by it, through the presence of type-3-advisers who are following the bonus and as whom such type-1s do not want to be perceived. However, while theoretically interesting, such behavior is of a very indirect nature and requires high level of strategic sophistication and foresight. We do not observe behavior which is in line with such reasoning.[19]

Incorporating image concerns for current actions would also increase the costs of initially recommending an option from $\mathcal{B}$ for those who do not think that it is advisable. In consequence, the share of type-3 relative to type-2-advisers would decrease. This means that there would be less

[19]First, we observe only a small minority of advisers who first recommended an option in $\mathcal{C} \setminus \mathcal{B} = \{B, C\}$ and then switched to recommend or chose for themselves an option in $\mathcal{B} = \{A\}$. Second, this minority is not in any meaningful way lower in NO BONUS than in BONUS (5.9% vs. 4.2% in table 5 and 3.9% vs. 4.2% in table 6) although the pattern of type-1s trying not to be perceived as type-3s is ultimately caused by the bonus.

initial recommendations for an option from $\mathcal{B}$ and less type-3-advisers for whom pooling on such a recommendation is attractive. However, the pooling creates the persistent effect. In addition, some advisers of the fourth type persistently recommend and choose from $\mathcal{C} \setminus \mathcal{B}$. Again, these predictions are not consistent with what we actually observe in our experiment – many advisers recommended an option from $\mathcal{B}$ and stuck to it. For all these reasons, we think that the present specification, where image concerns are with regard to past behavior, allows us to comprehensively demonstrate our main points in a realistic framework with predictions which match our findings.

References

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

Babcock, L., G. Loewenstein, S. Issacharoff, and C. Camerer (1995). Biased judgments of fairness in bargaining. *American Economic Review 85*(5), 1337–1343.

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

Bénabou, 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 124*(2), 305–370.

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*, 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.

Cain, D. M., G. Loewenstein, and D. A. Moore (2005). The Dirt on Coming Clean: Perverse Effects of Disclosing Conflicts of Interest. *Journal of Legal Studies 34*(1), 1–25.

Cain, D. M., G. Loewenstein, and D. A. Moore (2011). When Sunlight Fails to Disinfect: Understanding the Perverse Effects of Disclosing Conflicts of Interest. *Journal of Consumer Research 37*(5), 836–857.

CEA (2015). The Effects of Conflicted Investment Advice on Retirement. *Report by the White House's Council of Economic Advisers*.

Christoffersen, S. E. K., R. Evans, and D. K. Musto (2013). What Do Consumers' Fund Flows Maximize? Evidence from Their Brokers' Incentives. *Journal of Finance 68*(1), 201–235.

Cohn, A., E. Fehr, and M. A. Maréchal (2014). Business culture and dishonesty in the banking industry. *Nature 7529*(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 83*(2), 587–628.

Eyster, E. (2002). Rationalizing the Past: A Taste for Consistency. *Nuffield College Mimeograph*.

Falk, A. (2017). Facing Yourself: A Note on Self-Image. *HCEO Working Paper 2017-09*.

Falk, A. and J. Tirole (2016). Narratives, Imperatives and Moral Reasoning. *mimeo*.

Falk, A. and F. Zimmermann (2017a). Consistency as a Signal of Skills. *Management Science 63*(7), 2197–2210.

Falk, A. and F. Zimmermann (2017b). Information Processing and Commitment. *Economic Journal*, forthcoming.

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

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

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.

Gesche, T. (2016). De-biasing strategic communication? *University of Zurich Department of Economics Working Paper Series 216*(216).

Gino, F., M. I. Norton, and R. A. Weber (2017). Motivated Bayesians : Feeling moral while acting egoistically. *Journal of Economic Perspectives 30*(3), 189–212.

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, Identitiy, and Unethical Behavior. *mimeo*.

Grossman, Z. and J. van der Weele (2017). Self-Image and Willful Ignorance in Social Decisions. *Journal of the European Economic Association 15*(1), 173–217.

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.

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.

Inderst, R. and M. Ottaviani (2012). Competition through Commissions and Kickbacks. *American Economic Review 102*(2), 780–809.

Koch, C. and C. Schmidt (2010). Disclosing conflicts of interest - Do experience and reputation matter? *Accounting, Organizations and Society 35*(1), 95–107.

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

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

Li, M. and K. Madarasz (2008). When mandatory disclosure hurts: Expert advice and conflicting interests. *Journal of Economic Theory 139*, 47–74.

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

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

Loewenstein, G., C. R. Sunstein, and R. Golman (2014). Disclosure: Psychology Changes Everything. *Annual Review of Economics 6*, 391–419.

Malmendier, U. and K. Schmidt (2017). You Owe Me. *American Economic Review 107*(2), 493–526.

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.

Mijović-Prelec, D. and D. Prelec (2010). Self-deception as self-signalling: a model and experimental evidence. *Philosophical Transactions of the Royal Society B 365*, 227–40.

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.

Schwardmann, P. and J. van der Weele (2016). Decepetion and Self-deception. *mimeo*.

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

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

Appendix B – Proofs for Appendix A

Proof of Lemma 1

Let j be the joint p.d.f. associated with the joint c.d.f. J for (κ, λ). Accordingly, it holds that

$$K(x) = \int_0^x \int_0^\infty j(\kappa, \lambda) d\lambda d\kappa.$$

Full support for the joint p.d.f. j, i.e. $j(\kappa, \lambda) > 0$ for all $(\kappa, \lambda) \in \mathbb{R}_0^+ \times \mathbb{R}_0^+$, implies

$$K'(x) = \int_0^\infty j(x, \lambda) d\lambda > 0.$$

Repeating this for Λ proves part a). Part b) can be proven analogously, as for any $x > 0$

$$\begin{aligned} & \Lambda(y \mid x) = \Pr[\lambda \leq y \mid \kappa \leq x] = \left(\int_0^y \int_0^x j(\kappa, \lambda) d\kappa d\lambda \right) \Big/ \left(\int_0^\infty \int_0^x j(\kappa, \lambda) d\kappa d\lambda \right) \\ \Rightarrow \quad & \Lambda'(y \mid x) = \left(\int_0^x j(\kappa, y) d\kappa \right) \Big/ \left(\int_0^\infty \int_0^x j(\kappa, \lambda) d\kappa d\lambda \right) > 0 \end{aligned}$$

For part c), rewrite the conditions $\kappa \in [0, x]$ and $\kappa/\lambda \leq z$ by $\kappa \in \{0, \min\{x, z\lambda\}\}$. We can then write the conditional c.d.f. $R(z \mid x, y) = \Pr[\kappa/\lambda \leq z \mid \kappa \leq x, \lambda \geq y]$ with $x > 0$ as

$$R(z \mid x, y) = \left(\int_y^\infty \int_0^{\min\{x, z\lambda\}} j(\kappa, \lambda) d\kappa d\lambda \right) \Big/ \left(\int_y^\infty \int_0^x j(\kappa, \lambda) d\kappa d\lambda \right)$$

If $x > z\lambda$, the partial derivative of the above w.r.t. z is then given by

$$R'(z \mid x, y) = \left(\int_y^\infty \lambda \cdot j(z\lambda, \lambda) d\lambda \right) \Big/ \left(\int_y^\infty \int_0^x j(\kappa, \lambda) d\kappa d\lambda \right)$$

and strictly positive by full support of j. If $x \leq z\lambda$, this derivative is zero so that $R'(z \mid x, y) \geq 0$ always holds. □

Proof of Proposition 1

First note that for type-2 advisers, $r_1 \notin \mathcal{B}$. Since type-3-advisers have $r_1 \in \mathcal{B}$, $\Pr[\theta = 3 \mid r_1 \notin \mathcal{B}, o] = 0$. Advisers of type $\theta = 2$ therefore maximize $U(o \mid r_1 \notin \mathcal{B}, r^*) = u(v(o))$ by choosing $o^* = r^* = r_1 \notin \mathcal{B}$ for themselves. Advisers of type $\theta \in \{1, 3\}$ have recommended $r_1 \in \mathcal{B}$ such that they both can be inferred to be possibly of type $\theta = 3$. Suppose share τ_o of type-1-advisers choose for themselves such that $o \in \mathcal{B}$. Similarly, let π_o denote the share of type-3-advisers who choose for themselves $o \in \mathcal{B}$. The following posteriors then emerge:

$$\Pr[\theta = 3 \mid o \in \mathcal{B}, r_1 \in \mathcal{B}] = \frac{\pi_o \cdot \phi_3}{\tau_o \cdot \phi_1 + \pi_o \cdot \phi_3} \tag{3}$$

$$\Pr[\theta = 3 \mid o \notin \mathcal{B}, r_1 \in \mathcal{B}] = \frac{(1 - \pi_o) \cdot \phi_3}{(1 - \tau_o) \cdot \phi_1 + (1 - \pi_o) \cdot \phi_3} \tag{4}$$

It is easily verified that the latter posterior is weakly larger than the former if and only if $\tau_o \geq \pi_o$. If this condition applies, then it holds for type-1-advisers (for whom $o^* = o^*_{\mathcal{B}}$) that for any $o' \notin \mathcal{B}$

$$u(v(o')) - \lambda \cdot \Pr[\theta = 3 \mid o', r_1 \in \mathcal{B}] < u(v(o^*_{\mathcal{B}})) - \lambda \cdot \Pr[\theta = 3 \mid o^*_{\mathcal{B}}, r_1 \in \mathcal{B}].$$

If type-1-advisers chose $o' \notin \mathcal{B}$ they would suffer for two reasons: First, such choices are suboptimal in terms of maximizing their pecuniary utility $u(v(o))$. Second, choosing $o' \notin \mathcal{B}$ leads to a worse image utility through a higher probability to be perceived as type-3. Accordingly, in all equilibria with $\tau_o \geq \pi_o$ all type-1-advisers choose $o = o^* \in \mathcal{B}$. Therefore, $\tau_o = 1 \geq \pi_o$ has to hold for all equilibria in this class.

In the candidate equilibrium with $\tau_o = 1 \geq \pi_o$, all type-1-advisers choose $o = o^* = r^* = r_1 \in \mathcal{B}$. Type-3-advisers can thus pool with type-1s by choosing consistently from $\mathcal{B}$, i.e. $o = r_1 \in \mathcal{B}$, even though for them $o^* \notin \mathcal{B}$. They then choose their constrained optimum $o^*_{\mathcal{B}} \in \mathcal{B}$. If they do not choose consistently they can choose their preferred option $o^* \notin \mathcal{B}$ but instead reveal themselves as corruptible, i.e. as type-3-advisers. Using (3) and the assumption that in case of indifference they choose o^*, this means that type-3-advisers pool if the following holds:

$$u(v(o^*)) - \lambda < u(v(o^*_{\mathcal{B}})) - \lambda \cdot \frac{\pi_o \phi_3}{\phi_1 + \pi_o \phi_3} \quad \Leftrightarrow \quad \lambda > (u(v(o^*)) - u(v(o^*_{\mathcal{B}}))) \cdot \left(\frac{\phi_1 + \pi_o \phi_3}{\phi_1} \right)$$

Since for type-3-advisers $u(v(o^*)) > u(v(o^*_{\mathcal{B}})$, the threshold on the RHS of the second inequality grows in π_o, the share of type-3-advisers who choose $o \in \mathcal{B}$ to pool with type-1s. In addition, because they are type-3-advisers, $\kappa < u(b)$ has to hold. Therefore, the share π_o of pooling type-3-advisers has to solve

$$1 - \pi_o = \Lambda \left((u(v(o^*)) - u(v(o^*_{\mathcal{B}}))) \cdot \left(\frac{\phi_1 + \pi_o \phi_3}{\phi_1} \right) \,\middle|\, u(b) \right).$$

From Lemma 1 b), it follows immediately that both $\pi_o = 0$ and $\pi_o = 1$ cannot be solutions. Also, the above RHS is strictly increasing in π_o while its values are contained in the unit interval. The above LHS is simply the decreasing 45-degree-line over the unit square. Accordingly, there has to be a unique solution $\pi^*_o \in (0, 1)$.

Finally, we exclude other equilibria with $\tau_o < \pi_o$. In this case, the posterior (3) is strictly larger than (4). Since for type-3-advisers $o^* = o^*_{\mathcal{C} \setminus \mathcal{B}}$ holds, they then choose their preferred choice as

$$u(v(o^*)) - \lambda \cdot \Pr[\theta = 3 \mid o^*, r_1 \in \mathcal{B}] > u(v(o^*_{\mathcal{B}})) - \lambda \cdot \Pr[\theta = 3 \mid o^*_{\mathcal{B}}, r_1 \in \mathcal{B}].$$

Thus, all type-3-advisers choose $o \notin \mathcal{B}$ and reveal themselves. This corresponds to $\pi_o = 0$ and therefore contradicts an equilibrium with $\tau_o < \pi_o$. □

Proof of Proposition 2

Type-2-advisers have initially recommended $r_1 \notin \mathcal{B}$. As type-3-advisers have recommended $r_1 \in \mathcal{B}$, it holds that $\Pr[\theta = 3 \mid r_1 \notin \mathcal{B}, o, r_2] = 0$ and type-2s therefore maximize $U(o \mid r_1 \notin \mathcal{B}, r^*) = -\kappa \cdot 1[r_2 \neq r^*]$ by re-recommending $r_2 = r^* = r_1 \notin \mathcal{B}$.

First, consider the situation that advisable choices are projected from own choice ($r^* = o^*$). Then, share $1 - \pi^*_o \in (0, 1)$ of type-3-advisers have already revealed themselves as such by choosing $o \neq r_1$ in the own choice O (see Lemma 1). Therefore, their image concerns are invariant to r_2 as for them, $\Pr[\theta = 3 \mid o \neq r_1, r_2] = 1$ applies for every $r_2 \in \mathcal{C}$. They then maximize $U(r_2 \mid r_1, r_2, r^*) = -\kappa \cdot \mathbb{1}[r_2 \neq r^*] - \lambda \cdot \Pr[\theta = 3 \mid r_2, r_1, o]$ by recommending $r_2 = r^* \notin \mathcal{B}$. Therefore, they do not re-recommend their initial, biased recommendation $r_1 \in \mathcal{B}$.

Type-1-advisers and share $\pi_o \in (0, 1)$ of type 3-advisers who have not yet revealed themselves both look identical to an outside observer as both have a history of $o = r_1 \in \mathcal{B}$. Accordingly, hitherto unrevealed type-3-advisers can continue to pool with type-1-advisers. Denote with τ_{r_2} the share of type-1-advisers who recommend $r_2 \in \mathcal{B}$ and with π_{r_2} the share of type-3-advisers who recommend $r_2 \in \mathcal{B}$. This yields the following posteriors, conditional on not having previously revealed oneself (i.e. that $o = r_1$ holds):

$$\Pr[\theta = 3 \mid r_2 \in \mathcal{B}, r_1 \in \mathcal{B}] = \frac{\pi_{r_2} \cdot \pi^*_o \phi_3}{\tau_{r_2} \phi_1 + \pi_{r_2} \cdot \pi^*_o \phi_3} \tag{5}$$

$$\Pr[\theta = 3 \mid r_2 \notin \mathcal{B}, r_1 \in \mathcal{B}] = \frac{(1-\pi_{r_2}) \cdot \pi_o^* \phi_3}{(1-\tau_{r_2}) \cdot \phi_1 + (1-\pi_{r_2}) \cdot \pi_o^* \phi_3} \tag{6}$$

Posterior (6) is weakly larger than (5) if and only if $\tau_{r_2} \geq \pi_{r_2}$. If this condition holds, the payoff for type-1-advisers with $r^* \in \mathcal{B}$ from re-recommending r^* in R2 is always strictly larger than from recommending $r_2' \notin \mathcal{B}$, i.e.

$$-\lambda \cdot \Pr[\theta = 3 \mid r^*, r_1 \in \mathcal{B}] > -\kappa - \lambda \cdot \Pr[\theta = 3 \mid r_2', r_1 \in \mathcal{B}].$$

Thus, the only equilibrium with $\tau_{r_2} \geq \pi_{r_2}$ obeys $\tau_{r_2} = 1$. Hitherto unrevealed type-3-advisers who want to pool with type-1s have to choose analogously, i.e. $r_2 = r_1 \in \mathcal{B}$, even though this is not their advisable choice because for them, $r^* \notin \mathcal{B}$ holds. This allows them to not reveal themselves as corruptible so that their image costs λ are discounted by $\Pr[\theta = 3 \mid r_2 \in \mathcal{B}, o = r_1 \in \mathcal{B}]$. For this, they experience costs κ of recommending something not consider advisable. Plugging in the above posteriors with $\tau_{r_2} = 1$ yields

$$-\lambda < -\kappa - \lambda \cdot \frac{\pi_{r_2} \cdot \pi_o^* \phi_3}{\phi_1 + \pi_{r_2} \cdot \pi_o^* \phi_3} \quad \Leftrightarrow \quad \kappa < \lambda \cdot \frac{\phi_1}{\phi_1 + \pi_{r_2} \cdot \pi_o^* \phi_3}$$

as a condition for hitherto unrevealed type-3-advisers to continue pooling with type-1s. Note that Lemma 1b implies that for every κ multiplied with some factor, there is a mass of advisers with sufficiently high λ, i.e. with $\lambda > \kappa(\phi_1 + \pi_{r_2} \cdot \pi_o^* \phi_3)/\phi_1$. Therefore, $\pi_{r_2} = 0$ cannot be true. Also, the limit on κ/λ which the above inequality implies is only relevant to type-3-advisers (those advisers who have $\kappa < u(b)$) who have not revealed themselves in O (those with $\lambda < (u(v(o^*)) - u(v(o_{\mathcal{B}}^*))) \cdot (\phi_1 + \pi_o^* \phi_3)/\phi_1$). Thus, the share of type-3-advisers who continue to pool, denoted by π_{r_2}, is determined by the solution to

$$\pi_{r_2} = R\left(\frac{\phi_1}{\phi_1 + \pi_{r_2} \cdot \pi_o^* \phi_3} \,\middle|\, u(b), (u(v(o^*)) - u(v(o_{\mathcal{B}}^*))) \cdot \left(\frac{\phi_1 + \pi_o^* \phi_3}{\phi_1} \right) \right). \tag{7}$$

A solution $\pi_{r_2} = 0$ has been ruled out above. By Lemma 1c the above RHS is non-increasing in π_{r_2} and takes value in the unit interval. As the LHS is just the 45-degree line above it, there has to be a unique intersection for some $\pi_{r_2}^* \in (0, 1]$.

We exclude equilibria with $\tau_{r_2} < \pi_{r_2}$ in a similar fashion as in the proof of Lemma 1. With $\tau_{r_2} < \pi_{r_2}$, the posterior (5) is larger than (6). For hitherto unrevealed type-3-advisers with $r^* \notin \mathcal{B}$ it thus holds that

$$-\kappa - \lambda \cdot \Pr[\theta = 3 \mid r_2 \in \mathcal{B}, r_1 \in \mathcal{B}] < -\lambda \cdot \Pr[\theta = 3 \mid r^* \notin \mathcal{B}, r_1 \in \mathcal{B}]$$

and they all recommend the choice $r^* \notin \mathcal{B}$ and thereby reveal themselves. This implies $\pi_{r_2} = 0$ and thus contradicts an equilibrium with $\tau_{r_2} < \pi_{r_2}$.

Recall from Lemma 1 and its proof that when the recommendations which advisers consider advisable are independent from own choices, the own choice o has no diagnostic value and type-3-advisers have not had yet the possibility to reveal themselves in O. In terms of signaling value for R2, this is equivalent to the above when $\pi_o^* = 1$. The above reasoning can then be repeated with this parameter choice when the RHS in (7) is replaced by $R\left(\phi_1/(\phi_1 + \pi_{r_2}\phi_3) \mid u(b), 0\right)$ as no prior chance to reveal oneself does not restrict the subset of those type-3-advisers who can pool (i.e. it does not restrict the values of λ). The qualitative results, however, remain unchanged. □

Appendix C – Further statistics

All regression results in the main text are based on linear probability models estimated by OLS and 2SLS. In the following, we present the corresponding non-linear probit estimates for all regressions with a binary outcome variable. For single-equation (OLS) models, marginal effects based on probit estimates are presented. As the relevant independent variables are binary, these marginal effects are computed as the predicted average difference between the (conditional) expectation when the corresponding dummy is either zero or one. For the two-equation (2SLS) estimates in column 5 of table 2 and 4 a bivariate probit (biprobit) model was estimated and marginal effects were determined analogously, although their interpretation and presentation is different (see explanation of the †-symbol below).

Some combinations of the control variables predicted outcomes perfectly so the corresponding observations were dropped in the ML-based probit estimations. The corresponding smaller numbers of observation are denoted with a $^{\diamond}$-symbol. For the same reason and the resulting small number of degrees of freedom, the bivariate probit model did sometimes not converge. To circumvent this, we estimated these bivariate models with either personal *or* session controls; column 5a and 5b of table 9 and 10 thus correspond both to column 5 of table 3 and 4, respectively, in the main text.

	(1)	(2)	(3)
$Bonus$	0.520***	0.502***	0.543***
	(0.088)	(0.070)	(0.075)
Personal Controls	yes	no	yes
Session Controls	no	yes	yes
Observations	89$^{\diamond}$	99	89$^{\diamond}$

*Table 8: Average marginal effect based on probit estimates of the prob. to recommend option A in R1. Robust standard errors in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Personal controls: age, gender, monthly budget, subject's region of origin, degree and field of studies.*

	(1)	(2)	(3)	(4)	(5a)	(5b)
$Bonus$	0.213**	0.174**	0.216***	0.069		
	(0.091)	(0.071)	(0.070)	(0.079)		
$r_1 = A$				0.332***		
				(0.095)		
$\widehat{r_1 = A}$ (via $Bonus$)					0.558†	0.509†
					-	-
Personal Controls	yes	no	yes	yes	yes	no
Session Controls	no	yes	yes	yes	no	yes
Observations	79$^{\diamond}$	99	79$^{\diamond}$	66$^{\diamond}$	99	99

*Table 9: Average marginal effect based on (bi-)probit estimates of the prob. to choose option A in O. Robust standard errors in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Personal controls: age, gender, monthly budget, subject's region of origin, degree and field of studies.*

	(1)	(2)	(3)	(4)	(5a)	(5b)
$Bonus$	0.181**	0.222***	0.270***	0.118		
	(0.077)	(0.071)	(0.107)	(0.089)		
$r_1 = A$				0.357***		
				(0.127)		
$\widehat{r_1 = A}$ (via $Bonus$)					0.441†	0.507†
					-	-
Personal Controls	yes	no	yes	yes	yes	no
Session Controls	no	yes	yes	yes	no	yes
Observations	81◇	87◇	66◇	66◇	99	99

*Table 10: Average marginal effect based on (bi-)probit estimates of the prob. to recommend option A in R2. Robust standard errors in parentheses, significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Personal controls: age, gender, monthly budget, subject's region of origin, degree and field of studies.*

†-*explanation*: The 2SLS-estimate for the coefficient on $\widehat{r_1 = A}$ corresponds, in the limit, to the effect of having recommended option A for those who did so because of the bonus (i.e. the LATE-effect, see Chiburis et al., 2012[20]). The average marginal effect of $r_1 = A$ as estimated in a biprobit model corresponds to the mean difference in own choices or second recommendations between those who have initially recommended option A or not (i.e. the ITT-effect). Therefore, it does not take into account that this initial recommendation was due to the bonus. To make it comparable to the 2SLS(-LATE)-estimates, the biprobit marginal effect has to be divided by the marginal $Bonus$-effect on the initial recommendation in R1 (i.e. the estimated share of compliers) which was estimated simultaneously in the same biprobit-model. The corresponding estimates are portrayed here and denoted with the †-symbol. As these are the ratios of two estimated marginal effects, standard errors are not meaningful and omitted. Both of the underlying marginal effects are however significant at conventional levels.

The table on the next page presents subjects' summary statistics and the results of a randomization check across the BONUS-treatment and the baseline NO BONUS.

[20]Chiburis, R. C., J. Das, and M. Lokshin (2012). A practical comparison of the bivariate probit and linear IV estimators. *Economics Letters 117*, 762-766.

	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	0.021	0.144	0.010	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		

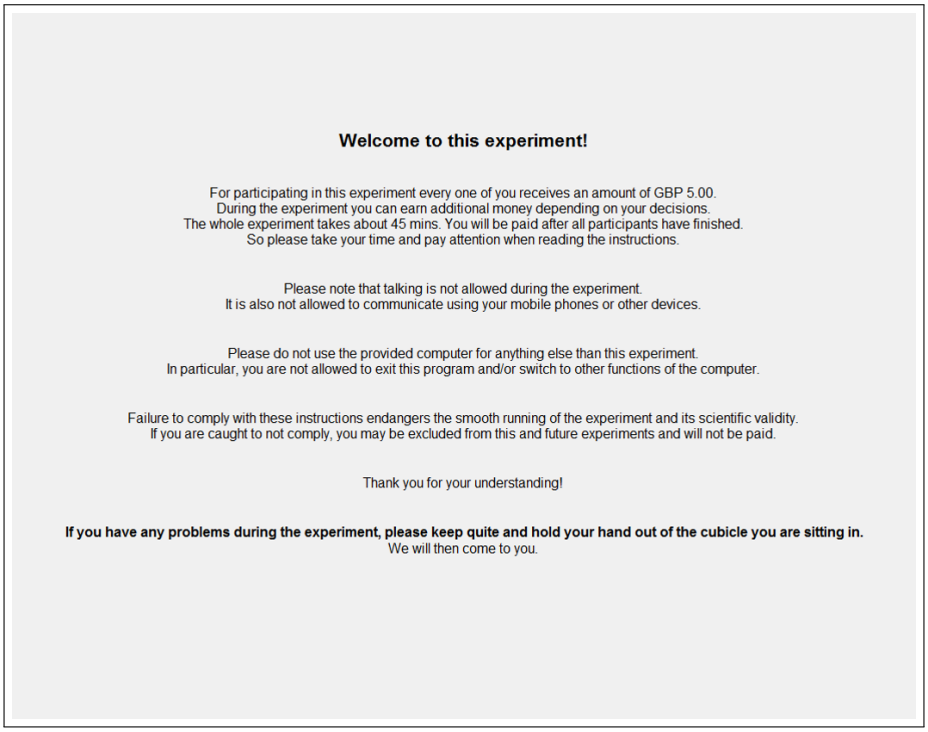
Table 11: Summary statistics for advisers' personal characteristics and dummy variable based on categorical data. The rightmost column provides p-values for a randomization check between NO BONUS and BONUS (Wilcoxon rank-sum tests for the variables age and budget; χ^2-tests for the remaining categorical variables).

Appendix D – Experimental instructions

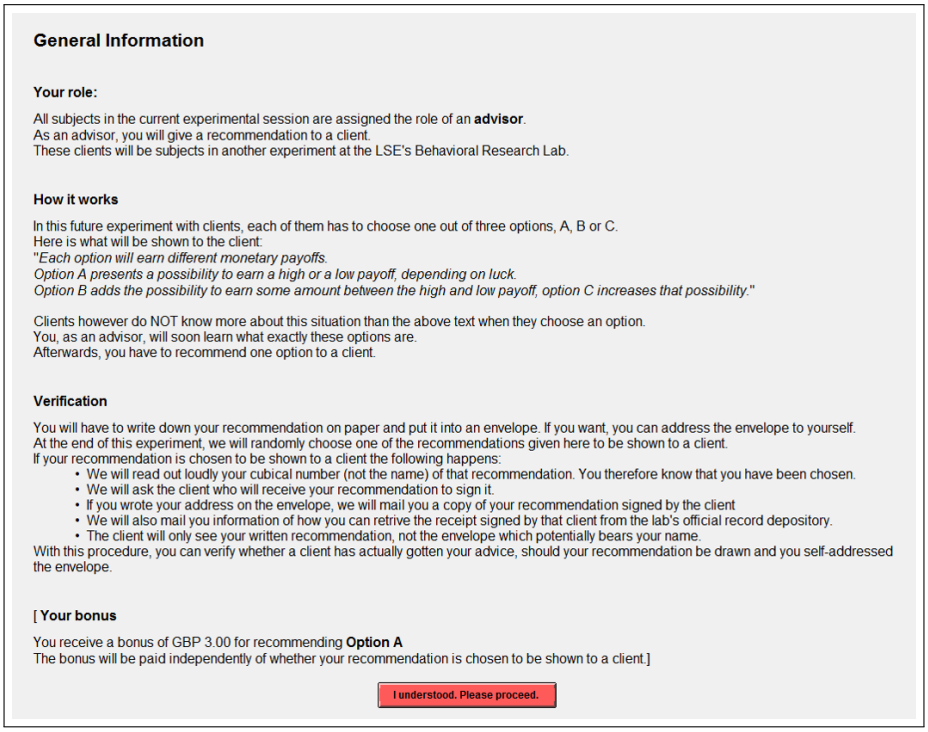
The following pages contain screenshots of instructions shown to subjects in ztree and on the information about the investment options printed on paper. They are presented in the order as they were seen by the subjects in the experiment.

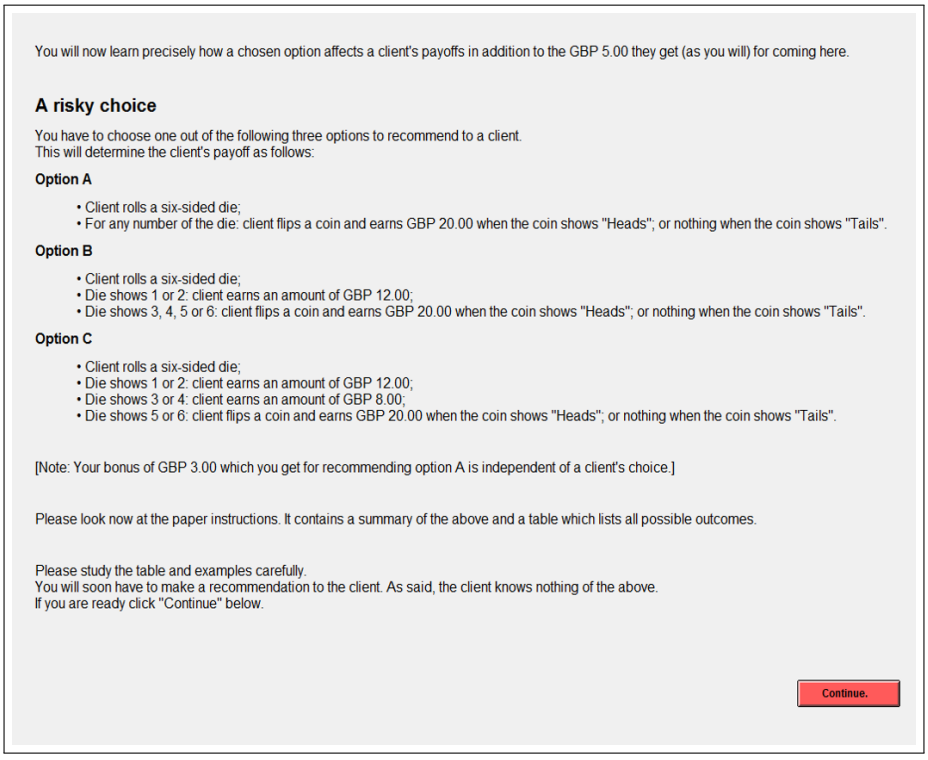
- Screen 1: Welcome stage and general instructions
- Screens 2a and 2b: Explanation for R1. Two screens which explain the client's choice situation, the adviser's role, and the investment options.
- Information on the investment options shown to advisers, printed on paper
- Screen 2c: Instructions for 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

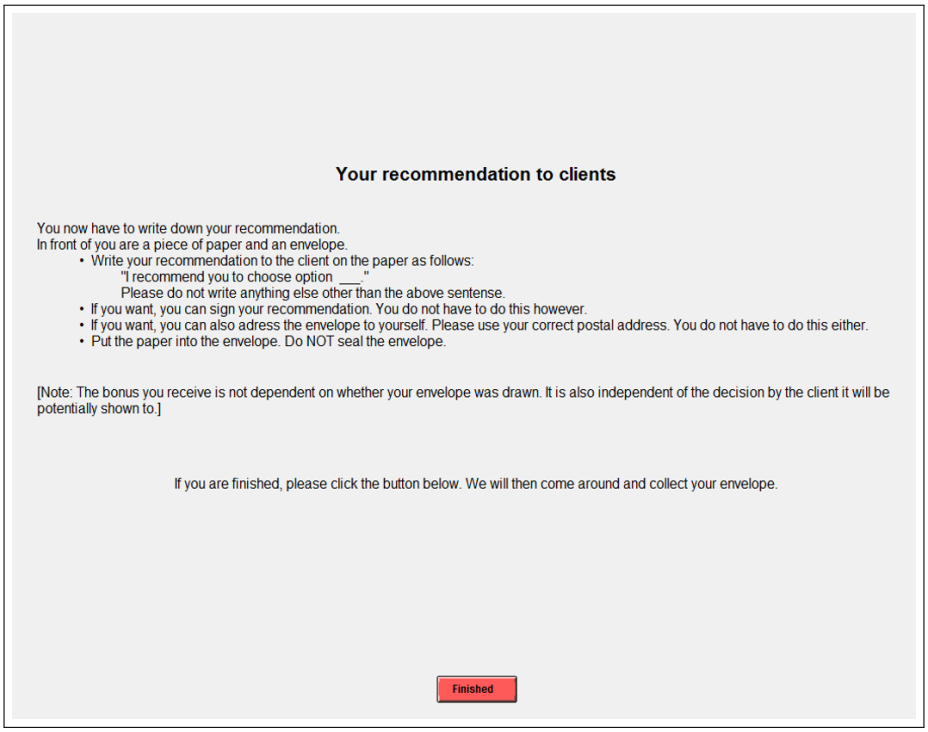
Example:

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.

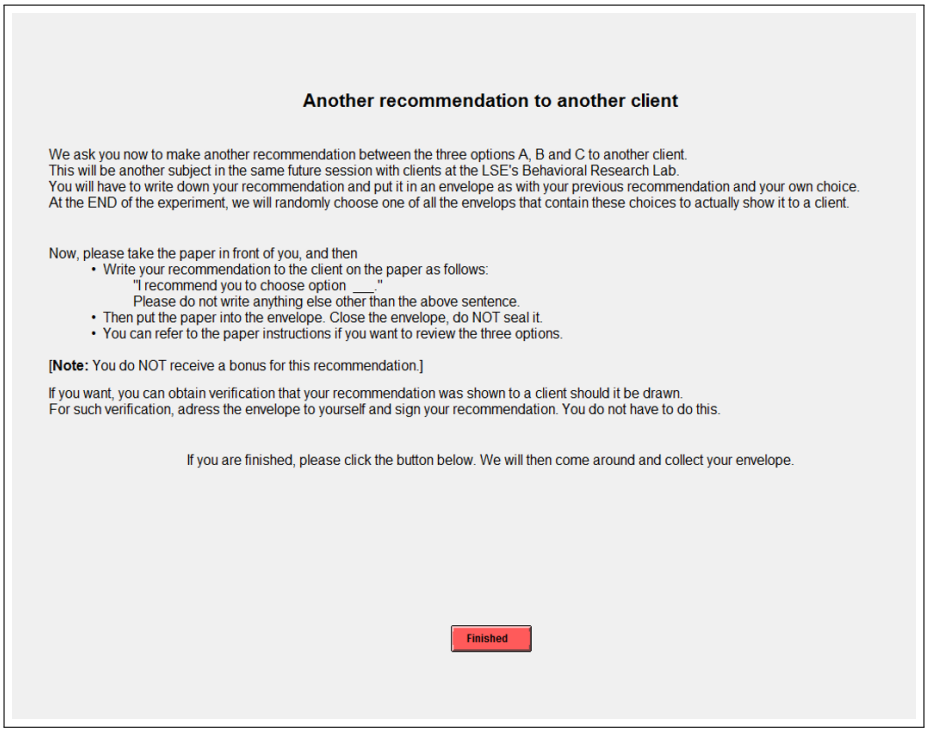
Suppose the die yielded 1. If option B or C was chosen before, one gets GBP 12.00 for sure. If option A was chosen, one plays the lottery instead.

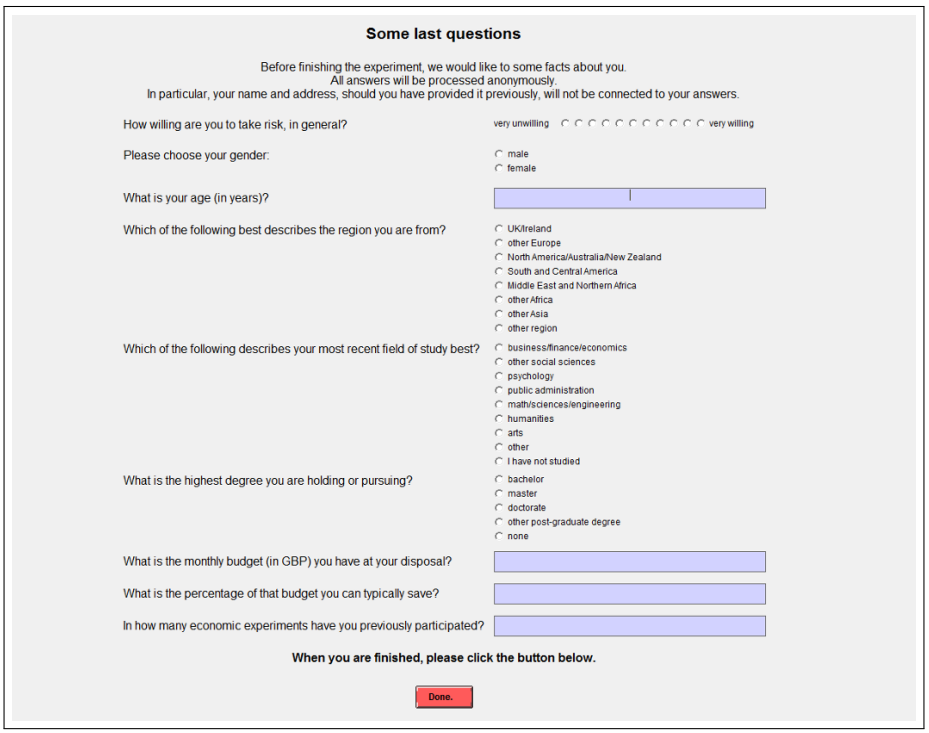
Suppose the die yielded 6. Independently of the chosen option one plays the lottery.

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)