Outcome bias and risk taking in a principal agent setting - experimental design and pre-analysis plan

Moritz Loewenfeld

March 4, 2025

1 Research Question

I consider a setting of delegated risk taking. Agents choose between a first-order stochastically dominant and a dominated lottery. Principals observe choices and outcomes of both lotteries and then decide whether to award a bonus payment to the agent. The goal of this experiment is to study whether outcome bias (OB), which is a tendency to condition bonus payments on outcomes, can shape the incentives faced by agents and thereby their choices. In particular, I seek to address the following research questions. 1) Can outcome bias in bonus decisions create incentives to choose suboptimal actions? 2) Do agents anticipate the OB of principals correctly and do they adjust their choices accordingly? I.e., can outcome bias induce more choices of sub-optimal actions and thus decrease welfare? I further seek to better understand the mechanisms driving outcome bias.

2 Experimental design

Agents make a number of decisions between two lotteries on behalf of the principals. Principals decide on bonus payments. Participants are randomly and permanently assigned to the role of either principal or agent.

There are two treatments. In the reward-after treatment, principals make bonus decisions for all possible choice-outcome combinations (strategy method). In the reward-before treatment, principals condition only on the agents' choices, but not their outcomes. Principals make a total of 42 bonus decisions in the reward-after treatment and a total of 14 bonus decisions in the reward-before treatment. See table 1 and 2 for the lottery tasks employed. In both treatments, the bonus amounts to 10 pounds. Whenever the bonus is not allocated to the agent matched with the principal, it is allocated to another randomly chosen agent.

The order in which principals make bonus decisions for the different choice or choice-outcome combination is randomized between subjects. Moreover, the display of the choice tasks (order of states, position of lotteries G and B) is randomized between scenarios (choice of the agent in reward-before treatment or choice and outcome in reward-after treatment) and subjects.

In addition to the bonus decisions, principals also make choices between the lotteries used in part I of the experiment for themselves.

I employ two main conditions, *robustness* and *consequences*. The goal of the robustness condition is mainly to replicate patterns found in previous work with a larger (online) sample. The purpose of the consequences condition is to create a setting in which the bonus decisions of a given principal can have a direct influence on their agent's choices.

Robustness condition: In the robustness condition, each agent makes 2 choices for each of the 7 lottery pairs. In addition, beliefs are elicited. Agents make a first choice for each of the 7 lottery pairs. Thereafter, their beliefs are elicited and they make a second choice for each lottery pair. In the reward-before treatment, agents are asked, for each choice task, how likely they are to receive the bonus when choosing either lottery in their choice set. In the reward-after treatment, agents are asked to state their beliefs conditional on their choice and the outcome of the lotteries. Beliefs are incentivized using the binarized scoring rule.

Consequences condition: Each agent is matched to one principal and makes choices for 4 randomly chosen choice tasks. Agents make 15 choices for each choice task. After each choice, they are informed whether or not they received the bonus.

Revisions: Principals in the robustness and the consequences condition are given the opportunity to revise some of their bonus decisions. For each principal, the computer randomly selects one lottery pair from pairs 1-3 (see table 2). For each of the two selected choice tasks, principals receive a summary of their reward decisions and are asked to revise them. If OB is a cognitive bias, principals might "correct" their bonus decisions when they are made aware of the resulting incentives. Agents in the robustness condition are similarly invited to revise two choices after they have received a summary of their stated beliefs.

Payments: Each Agent is randomly paired with one principal. For each pair, one of the actions taken by the agent is randomly chosen. The action of the agent and the bonus decision of the principal is implemented. With 80% probability, participants are paid based on the principal-agent interaction. With 20% probability, principals are paid based on their choices in the risk tasks, and agents are paid based on their beliefs (random-incentive mechanism).

Participants receive a participation fee of 1.75 pounds, and 10% will be randomly selected to have their choices paid out. The expected additional earnings are around 1 pound.

All subjects will further answer a questionnaire (see section 2.1).

Procedures: The experiment will be conducted on Prolific, in late March of 2025. The target number of participants is 110 valid submissions per role and treatment in the conditions robustness and consequences. In a first step, data from principals is selected. Once this collection is complete, data from agents will be collected. Each subject has to pass three captchas and two attention check. Subjects who fail two or more of these five checks are excluded from the analysis.

	Correlation 1					Correlation 2			
	1(1/3)	2(1/3)	3(1/3)			1(1/3)	2(1/3)	3(1/3)
G	H $+\epsilon$	M $+\epsilon$	L $+\epsilon$	$\rightarrow$	G	H $+\epsilon$	M $+\epsilon$	L $+\epsilon$
B	M	L	H		B	L	H	M

Table 1 The first row presents different possible states and their probability of occurring. Rows 2 and 3 display the payoffs of option G (FOSD) and B in the different states of the world.

lottery pair	H	M	L	ϵ
1	1953	1031	109	45
2	2000	900	10	80
3	1708	810	33	115
4 (corr 2 only)	1403	688	103	523

Table 2 Parameter values for the different lotteries. All payoffs are in pence.

2.1 Questionnaire items

The questionnaire (non-incentivized) will contain the following items:

- Three items from an extended version of the cognitive reflection test (Frederick 2005, Toplak et al. 2014)
- Standard demographics will be obtained through the prolific database.

Moreover, subjects in the role of principal are asked to what extend their bonus decisions were impacted by 1) the agent's choice 2) the obtained outcome 3) a comparison between the obtained and the forgone outcome 4) a tendency to award the bonus to the matched agent rather than to a randomly chosen agent. Agents are asked to what extend their lottery choices were driven by 1) a desire to make good choices 2) maximization of the probability to obtain the bonus and 3) whether they sometimes made choices they thought were not in the best interest of the principal because this might not maximize their reward probability.

3 Model

The experimental design and predictions are based on a model in which bonus decisions depend on counterfactual evaluation, that is, a comparison between the payoff the agent obtained with the forgone payoff, the payoff they could have obtained, had they chosen a different lottery. In the model, principals are motivated by reciprocity to reward agents for good choices. However, their perception of what the good choice is biased by observing the outcome.

The principal's utility is described by some value function $v()$. If state s materializes, the principal enjoys an (ex-post) utility $v(x_s^\theta)$ from the payoff x_s^θ yielded by lottery $\theta \in \{G, B\}$ in this state. The ex-ante value of lottery $\theta \in \{G, B\}$ is given by $V(\theta) = \sum_{s \in S} p_s v(x_s^\theta)$.[1]

In the model, if state s materialized, the principal rewards a choice of the lottery G if and only if

[1] $v()$ could be, but need not be, a v.N.M utility function in which case the decision maker's preferences would satisfy expected utility theory.

$$\Delta(\tilde{V}_s) = \lambda \underbrace{[v(x_s^G) - v(x_s^B)]}_{\text{Ex-post comparison}} + (1-\lambda) \underbrace{[V(G) - V(B)]}_{\text{Ex-ante comparison}} \geq 0$$

,where $\lambda \in [0,1]$ denotes the principal's degree of outcome bias. If the agent chose lottery B, the principal's perceived goodness of the agent's choice is $-\Delta(\tilde{V}_s)$. The key features of the model are that the bonus probability is increasing in the quality of choice (choosing the FOSD lottery), increasing in the obtained outcome and decreasing in the forgone payoff.

Denote $P_{lp,j}(\theta^{corr}) = \sum_s p_s P_{j,lp,s}(\theta^{corr})$ the probability of receiving the reward for lottery pair $lp \in \{1,2,3\}$, from a principal $j \in J$, after lottery choice $\theta \in \{G,B\}$ under correlation structure $corr \in \{1,2\}$.[2] The incentives for the agents to choose option G rather than B for a given lottery pair lp under correlation structure $corr \in \{1,2\}$ set by principal j are given by the difference in probabilities of receiving the bonus when choosing option G instead of option B, that is $I_{lp,j}(G^{corr}) = P_{lp,j}(G^{corr}) - P_{lp,j}(B^{corr})$.

The key implication of the model is that outcome bias can create perverse incentives, that is incentives that contradict the principal's own revealed preferences. Under correlation 2, there exists $\overline{\lambda}_{lp}$ such that $I_{lp,j}(G^2) < 0$ for all $\lambda_j \in (\overline{\lambda}_{lp}, 1]$. Crucially, this can be the case even if a principal has a revealed preference for lottery G, in the sense that they would choose G if they had to choose for themselves. However, even for $\lambda_j = 1$, agents have positive incentives to choose the dominant lottery under correlation 1, i.e. $I_{lp,j}(G^1) > 0$ for all $lp \in \{1,2,3\}$. This implies that perverse incentives are hypothesized under correlation 2 but not 1.

Within the model, the reward-before treatment can be thought of as forcing $\lambda = 0$. The reward-before condition thus provides a baseline against which to compare the behavior in the reward-after treatment. The above discussed patterns should not occur in this treatment.

4 Main Hypotheses

All null hypotheses in this section will be tested at the 5% significance level. Whenever hypotheses are tested with logistic regressions, I employ standard Wald Chi-Square tests. All main hypotheses are tested on bonus decisions and choices that where made prior to a revision, or result from choices of agents matched to non-revised bonus decisions.

4.1 Variable construction

Denote $P_{lp,j}(\theta^{corr}, T)$ the probability of receiving the reward for lottery pair $lp \in \{1,2,3\}$ when being paired with principal j, after choice $\theta \in \{G,B\}$ and in treatment $T \in \{after, before\}$. In the $before$ treatment, $P_{lp,j}(C^{corr}, before) \in \{0,1\}$. The probability of receiving the bonus after choosing a given option in the reward-after treatment will be calculated as $P_{lp,j}(C^{corr}, after) = \sum_s p_s P_{lp,j,s}(C^{corr}, after)$, where s are the possible states of the world and p_s the associated prob-

[2] The lottery for $lp = 5$ is included solely to aid the structural estimation of λ and is therefore not included in the discussion.

abilities. Hence $P_{lp,j}(C^{corr}, before) \in \{0, 1/3, 2/3, 1\}$. Lottery pair 4, correlation 2, is mainly included for structural estimation and will not be used when testing the main hypotheses.

I define the incentives to choose option G rather than B as the difference in probabilities of receiving the bonus when choosing option G instead of option B, that is $I_{lp}(G^1, T) = P_{lp,j}(G^{corr}, T) - P_{lp,j}(B^{corr}, T)$. In all ensuing tests, I will omit the subscripts j and lp for brevity.

4.2 Preliminary: Outcome bias in bonus decisions

Hypothesis 1. ***Outcome bias in bonus decisions:*** *Principals are more likely to award the bonus if*

a) Agents chose their preferred lottery.

b) The obtained outcome is greater than the forgone outcome.

I estimate the following random-effects logit regression model.

$$Bonus_{i,t} = \beta_0 + \beta_1 preferred_{i,t} + \beta_2 obtained\ payoff_{i,t} + \beta_3 \{obtained > forgone\}_{i,t} + \epsilon_{i,t} \quad (1)$$

, where $Bonus_{i,t}$ is a dummy variable indicating whether the principal awarded the bonus or not, $preferred_{i,t}$ indicates whether the agent chose the principal's preferred lottery, as measured by her own choices, $obtained\ payoff_{i,t}$ denotes the payoff the principal obtained, in Euro, and $\mathbb{1}\{obtained > forgone\}_{i,t}$ is a dummy variable indicating whether the obtained outcome is higher than the forgone alternative. The regression model will be estimated for the reward-before and the reward-after condition separately.

To test hypothesis 1, I test the following null hypotheses using Wald Chi-Square tests, with standard errors clustered at the subject-level.

- For hypothesis 1a) I test the null hypothesis that $\beta_1 = 0$, against the alternative hypothesis that $\beta_1 > 0$, for both treatments
- For hypothesis 1b) In the reward-after (reward-before) treatment I test the null hypothesis that $\beta_3 = 0$, against the alternative hypothesis that $\beta_3 > 0$ ($\beta_3 \neq 0$). For the reward-before treatment, this hypothesis is not expected to be rejected ("placebo-test").[3]

4.3 OB and the creation of perverse incentives

The main hypothesis to be tested is that outcome bias can create perverse incentives.

Hypothesis 2. ***OB creates perverse incentives:***

a) Principals in the reward-after condition will be more likely to incentivize agents to choose the dominated lottery although they display revealed preferences for the dominant lottery under correlation 2 than under correlation 1, or their counterparts in the reward-before treatment for either correlation structure.

[3] For clarity: Whenever a clear direction of the alternative is stated ($<$, or $>$, one sided tests will be used. If no direction is stated for the alternative ($\neq$), two-sided tests will be used.

To test this hypothesis, I create a dummy $PI_{i,lp}$, that equals 1 if a given principals displays a revealed preference for lottery $G(B)$ but their bonus decisions imply strict incentives to choose lottery B (G), where incentives are as defined above. To formally test whether outcome bias can create perverse incentives, I will run the following random-effects logistic regression.

$$PI_{i,lp} = \beta_0 + \beta_1 after_{i,lp} + \beta_2 corr2_{i,lp} + \beta_3 after_{i,lp} * corr2_{i,lp} + \epsilon_{i,t}. \tag{2}$$

$after_{i,lp}$ is a dummy that equals 1 if a principal was assigned to the reward-after condition, and $corr2_{i,lp}$ is a dummy that equals 1 if a lottery has correlation structure 2. The main hypothesis derived from the model is that $\beta_3 > 0$.

4.4 Aggregate level hypotheses on Incentives

In a previous study I have tested hypotheses regarding aggregate effects of OB on incentives. In hypothesis 3, I collect hypotheses on how OB affects aggregate incentives to choose the dominant lottery.

Hypothesis 3. ***Outcome bias and incentives-aggregate level:***

a) $I(G^1, before) > 0$ *and* $I(G^2, before) > 0$.

b) $I(G^{corr}, before) > I(G^{corr}, after)$*, for* $corr \in \{1, 2\}$.

c) $I(G^1, before) - I(G^2, before) < I(G^1, after) - I(G^2, after)$.

I will test the following null hypotheses.

- For hypothesis 3a) I test the null hypotheses that $P(G^{corr}, before) = P(B^{corr}, before)$, for $corr \in \{1, 2\}$, against the alternative hypothesis that $P(G^{corr}, before) > P(B^{corr}, before)$. Wilcoxon signed-rank tests will be used.
- For hypothesis 3b), I test the null hypotheses that $I(G^{corr}, before) = I(G^{corr}, after)$, for $corr \in \{1, 2\}$, against the alternative that $I(G^{corr}, before) > I(G^{corr}, after)$. Wilcoxon rank-sum tests will be used.
- For hypothesis 3c), I test the null hypotheses that $I(G^1, before) - I(G^2, before) = I(G^1, after) - I(G^2, after)$, for $corr \in \{1, 2\}$, against the alternative that $I(G^1, before) - I(G^2, before) < I(G^1, after) - I(G^2, after)$. A wilcoxon rank-sum test will be used.

4.5 Differences across conditions

I will test the hypotheses outlined above for the two conditions separately. The main purpose of the consequences condition is to test whether perverse incentives occur also in a setting where principals can directly impact their agent's choice with their reward decision. The expectation is that outcome bias and perverse incentives do still occur, though they might be attenuated.

To test whether OB is attenuated in the consequences condition with respect to the robustness condition, I run logistic regressions similar to the one in (1):

$$Bonus_{i,t} = \beta_0 + \beta_1 preferred_{i,t} + \beta_2 obtained\ payoff_{i,t} + \beta_3 \{obtained > forgone\}_{i,t} + \tag{3}$$

$$\beta_4 consequences_{i,t} + consequences_{i,t}[\beta_5 preferred_{i,t} + \beta_6 obtained\ payoff_{i,t} + \beta_7 \{obtained > forgone\}_{i,t}] + \tag{4}$$

$$\epsilon_{i,t} \tag{5}$$

, where $consequences_{i,t}$ is a dummy that equals 1 if a subject was in the consequences treatment. For this regression, I pool observations for all lottery pairs and the robustness and consequences condition. I run this regression separately for the reward-before and the reward-after treatment. In the reward after treatment, a shift towards rewarding based on choices due to consequences would imply $\beta_5 > 0$, $\beta_6 < 0$, and $\beta_7 < 0$. I will test these hypotheses using two-sided tests.

4.6 The agent's choices

Denote $F(G^{corr}, treatment)$ the frequency with which agents choose the dominant lottery under correlation $corr \in \{1, 2\}$, in $treatment \in \{before, after\}$.

Hypothesis 4. ***The agents' choices***

a) In the reward-before treatment, agents choose the dominant lottery at a high frequency, under both correlation structures. Moreover, $F(G^{corr}, before) > F(G^{corr}, after)$*, for* $corr \in \{1, 2\}$.

b) $F(G^1, before) - F(G^2, before) < F(G^1, after) - F(G^2, after)$.

I will test these hypotheses running the following random-effects logistic regression, pooling all lottery pairs, for the robustness and the consequences condition separately. For the robustness treatment, I will include both choices the agents make. For the consequences condition, I will include all choices an agent makes for a given lottery.

$$G_{i,t} = \beta_0 + \beta_1 after_{i,lp} + \beta_2 corr2_{i,lp} + \beta_3 after_{i,lp} * corr2_{i,lp} + \epsilon_{i,t}. \tag{6}$$

$G_{i,t}$ is a dummy that indicates whether an agent chose the dominant lottery.

- For hypothesis 4a) I test the null hypotheses that $\beta_1 = 0$ against the alternative that $\beta_1 < 0$.
- For hypothesis 4b) I test the null hypotheses that $\beta_3 = 0$, against the alternative that $\beta_3 < 0$.

5 Revisions

5.1 Principals

Under the assumption that conditioning bonus decisions on an ex-post comparison of outcomes results from a cognitive bias, i.e. is a "mistake", principals should correct their bonus decisions once they receive a summary of their decisions and the resulting incentives faced by agents. If this error correction hypothesis holds, principals should revise more of their bonus decisions under correlation 2 than 1, and this tendency should be more pronounced in the reward-after than in the reward-before condition. Alternatively, if the conditioning of bonus decisions on an ex-post comparison of outcomes is deliberate, no corrections might occur, or principals might correct in the direction of more OB.

On the aggregate level, I test the hypothesis of error correction as follows. I define a variable $revise_{i,t}$ that equals 1 if a principal changes a given bonus decision and 0 otherwise. I run the following random-effects logistic regression

$$revise_{i,t} = \beta_0 + \beta_1 after_{i,lp} + \beta_2 corr2_{i,lp} + \beta_3 after_{i,lp} * corr2_{i,lp} + \epsilon_{i,t}. \tag{7}$$

where $after_{i,lp}$ and $corr2_{i,lp}$ are defined as above. I test the null that $\beta_3 = 0$ against the alternative hypothesis that $\beta_3 > 0$.

It is of course possible that different principals revise their bonus decisions in different ways. I will explore this possibility using clustering analysis such as latent class analysis or kmeans clustering.

5.2 Agents

In a previous experiment, I found that many agents anticipate incentives to choose the dominated lottery for correlation structure 2 (according to stated beliefs) but choose the dominant lottery. If this inconsistency between actions and stated beliefs is caused by a failure of strategic reasoning, agents might revise their choices when receiving a summary of their stated beliefs and the implied (expected) incentives. Assuming that agents will display similar behavior in this experiment (prior to the revision), agents should be more likely to revise their decisions under correlation structure 2 than 1, and this tendency should be more pronounced in the reward-after than in the reward-before condition. I test this hypothesis using a regression approach similar to that specified in equation 7. Again, the hypothesis to be tested is that $\beta_3 > 0$.

6 Further analysis

In a preliminary step, I will test for correlation sensitivity in the principals' lottery choices they make for themselves. The hypotheses above are derived under the assumption that the change in

the correlation structure does not meaningfully influence the principals' preferences.

An important question is to what extent agents are capable of anticipating the principals' OB and how the OB impacts their incentives to choose between the different lotteries. I will thus analyze how well agents' beliefs reflect the principals' bonus decisions. Although the working hypothesis is that agents form accurate beliefs, this analysis is descriptive and somewhat exploratory in nature. Therefore, no specific hypotheses are specified here.

I will further estimate the model structurally. For each principal in the reward-after condition, I will estimate an individual level of outcome bias λ. From agents' beliefs in the reward-after condition, I will estimate their perceived level of outcome bias in the population of principals. This exercise will allow quantifying OB and perceived OB within my model. It will also facilitate the study of heterogeneity. In particular, I will test for a correlation between the principals' λ and their performance in the extended CRT. Data from a previous experiment indicated that higher values of λ are correlated with lower CRT scores, which motivates examining this particular correlation.

I will further explore correlations between demographic variables, questionnaire responses, and subjects' behavior.

References

Frederick, S. (2005), 'Cognitive reflection and decision making', *Journal of Economic perspectives* **19**(4), 25–42.

Toplak, M. E., West, R. F. & Stanovich, K. E. (2014), 'Assessing miserly information processing: An expansion of the cognitive reflection test', *Thinking & reasoning* **20**(2), 147–168.