THÈSE de DOCTORAT

de l'UNIVERSITÉ TOULOUSE CAPITOLE

Présentée et soutenue par

Monsieur Moritz LOEWENFELD

Le 7 juin 2024

Essais sur la sensibilité à la corrélation dans la prise de décision économique - une exploration théorique et expérimentale

École doctorale : **TSE Toulouse Sciences Economiques**
Spécialité : **Sciences Economiques - Toulouse**
Unité de recherche : **TSE-R - Toulouse School of Economics - Recherche**

Thèse dirigée par *Madame Astrid HOPFENSITZ*

Composition du jury
Rapporteur : M. Aurélien BAILLON
Rapporteur : M. Ferdinand VIEIDER
Examinateur : M. Sébastien POUGET
Directrice de thèse : Mme Astrid HOPFENSITZ

Toulouse School of Economics

Essays on correlation-sensitivity in economic decision-making

Ph.D. Thesis

Moritz Loewenfeld

Toulouse School of Economics

April 8, 2024

Acknowledgements

My PhD has been a bumpy road at times. It has posed many challenges to overcome. Some were the expected trials and tribulations that come with a PhD, others came as utterly unwanted surprises. Somehow, I made it to the other end and learned one or two valuable lessons along the way, about research but also about life in general. Along my path, I've greatly benefited from the insights, experience, support, and friendship of many, without whom I would not be where I am today.

I would like to start by thanking Astrid Hopfensitz for being my thesis supervisor. Throughout my PhD, Astrid has been a continued source of invaluable advice and support. I learned a great deal about the science and art of running experiments from her. When we finally ran our joint experiment in together at GATE, it really felt like a landmark in my PhD, a moment of passing on the torch, and a special event in our mentor-mentee relationship.

I'd also like to thank Sébastien Pouget for invaluable guidance, especially on my job market paper. His exceptionally high standards were tough to meet at times, but with his kind and supportive yet insisting way, he pushed me to stop nowhere short of excellence. I also owe him my visit to the Economic Science Institute at Chapman University. I have very fond memories of our three-hour meetings that took us from the office to lunch, to strolls under the Californian sun, and left my brains completely fried.

I'm also very grateful to Jiakun Zheng. Over the years, he has become a great friend and coauthor. It was with Jiakun that I started to discover regret theory and correlation sensitivity more broadly. Jiakun has been not only the best coauthor I could hope for but also an awesome mentor and teacher from whom I learned a lot through our joint projects. As our ideas for joint ventures are multiplying at a rate faster than we can keep pace with, I look forward to the many more existing projects to come.

A number of other researchers have played a significant role during my PhD. I would like to thank Ingela Alger for her continued support, valuable feedback on my research at many stages, her help with administrative issues relating to experiments, and her firm commitment to the behavioral group at TSE. I thank Sophie Moinas for her always helpful comments and suggestions on my research, especially the structural estimation in my job market paper, and for great feedback on my job market talk. I would also like to thank Nour Medahhi for his work as director of the doctoral program. I'm particularly grateful for his firm support and for giving me the kind of advice you need to hear but do not necessarily want to hear. Aurélien Baillon might have been the most time-efficient in making a difference in my PhD. Half an hour of discussion led me to greatly

sharpen large parts of the paper that constitute chapter 1. I thank Daniel Chen for the fun and instructive encounters with empirical research and for helping me with the funding for my sixth year. I further thank Stéphane Cezera for his help carrying out my experiments and Ulrich Hege for his great work as placement officer. I have benefited from thoughtful feedback from many at TSE, including Maximilian Müller, Andrew Rhodes, Jacques Crémer, Karine Van Der Straeten, Tiziana Assenza, Roberta Dessi, Christophe Bisière, and Jean Tirole.

I would also like to thank Stephen Ressenti for inviting me to spend a semester at the Economic Science Institute at Chapman University. I further thank David Rojo Arjona for stimulating discussions and fun soccer matches. I'd also like to thank Aaron Berman, Ginger Liu, and Cindy for making my stay at the ESI a very enjoyable and memorable experience.

I further thank my friends with whom I've shared many experiences, good and bad, during the PhD. I thank my office mates Esteban Muñoz Sobrado, Pau Juan Bartroli, Enrico Mattia Salonia, Alice Hallman, Antoine Jacquet, Pedram Pooyafar, and Lony Bessagnet. As a behavioral/experimental economist, it is easy to sometimes feel like somewhat of an outsider in TSE. Sharing this lot with you more than made up for the downsides. Thanks for all the interesting discussions and feedback on research projects, the innumerable lunches at CROUS, shared beers, and having the heart and guts to come and listen to me talk about outcome bias for the umpteenth time. A special thanks to Esteban for the many coffee breaks and trips to the supermarket for much-needed fuel during the job market process.

I'm particularly thankful to Gökçe Gökkoca for being the most amazing, kind, and thoughtful friend and roommate one could ask for. I thank Lisa Botbol for being her radiant and unadulterated self. I thank Tuuli Vanhapelto, Tim Ederer, Anaïs Fabre, Max Lesellier, Hippolyte Boucher, Léa Bignon, Peter Neis, José Alfonso Muñoz-Alvarado, Gosia Majewska for our many joint adventures including hikes, climbing trips, exploring Aveyron, and a much-needed stop at McDonald's.

Finally, I thank my family, especially Heidi, Christoph, Kai, Christine, Tobi, and Michi, for their loving support and understanding of the eccentricities of my academic life - and for occasionally reminding me that there is a world outside of academia. For this, and simply for being my friends, I also thank Mark, Salome, Paul, and Fintan.

Summary

This thesis is dedicated to studying correlation sensitivity in decision-making under risk. Allowing risk preferences to be sensitive to the correlation between lottery outcomes can explain classical deviations from expected utility theory as well as phenomena in various real-world settings. However, experimental evidence on correlation sensitivity is limited and mixed. Moreover, the concept of correlation sensitivity has thus far been studied almost exclusively in the context of individual decision-making. Chapters 1 and 2 seek to contribute to our understanding of correlation-sensitive risk preferences, whereas Chapter 3 explores how choices can become correlation-sensitive in a setting of delegated decision-making, even when none of the involved parties has correlation-sensitive preferences.

In the first chapter, together with Jiakun Zheng, we study correlation-sensitive preferences in individual decision-making. We show that correlation-sensitive preferences in the general framework of (Lanzani, 2022) can be classified into three categories. We propose a choice task to classify experimental subjects accordingly. In multiple experiments, we find that aggregate choices display correlation sensitivity but in the opposite direction as assumed in regret and salience theory. Clustering analysis identifies a consistently correlation-sensitive minority driving aggregate patterns, with the majority showing no correlation sensitivity. Crucially, the analysis does not produce a regret/salience theory type. We disentangle correlation sensitivity arising from deliberate within-state comparisons from incidental payoff comparisons due to the framing of decision problems. Both produce correlation sensitivity, with deliberate comparisons exerting a somewhat greater influence.

In the second chapter, also together with Jiakun Zheng, we reconsider recent experiments that seem to show evidence for correlation sensitivity as assumed in salience theory. However, these studies fail to control for event-splitting effects (ESE). We seek to disentangle the role of correlation and event-splitting in two settings: 1) the common consequence Allais paradox as studied by Bordalo et al. (2012), Frydman and Mormann (2018), Bruhin et al. (2022) 2) choices between Mao pairs as studied by Dertwinkel-Kalt and Köster (2019). In both settings, we find evidence suggesting that recent findings supporting correlation effects are largely driven by ESE. Once controlling for ESE, we find no consistent evidence of correlation effects. We conclude that our results thus shed doubt on the validity of salience theory in describing risky behavior.

In the third chapter, I leave the realm of individual decision-making and consider a setting of delegated decision-making. In a theory-guided experiment, I study how outcome bias, a tendency of principals to reward and punish economic agents as if they could have anticipated a random

state of the world, shapes the incentives and choices of agents. Agents choose between two lotteries on behalf of their principal. One lottery is first-order stochastically dominant, but the dominated lottery is more likely to yield a higher payoff state-by-state. Despite perfectly observing the agent's choice, principals tend to reward agents if they choose the lottery, which realizes a higher payoff. As a result, they incentivize agents to choose the dominated lottery. Although most agents anticipate these incentives, only strategically sophisticated agents tend to choose the dominated lottery when they believe they have an incentive to do so. Structural estimation suggests that principals are either fully outcome-biased or fully unbiased, with less cognitively sophisticated principals displaying more outcome bias. The results imply that outcome bias might be most relevant in settings where sophisticated agents meet relatively unsophisticated principals.

Resumé

Ce mémoire est consacré à l'étude de la sensibilité à la corrélation dans la prise de décision sous risque.

Dans le premier chapitre, en collaboration avec Jiakun Zheng, nous étudions les préférences sensibles à la corrélation dans la prise de décision individuelle. Nous montrons que les préférences sensibles à la corrélation dans le cadre général de (Lanzani, 2022) peuvent être classées en trois catégories. Nous proposons une tâche de choix pour classer les sujets expérimentaux en conséquence. Dans plusieurs expériences, nous constatons que les choix agrégés montrent une sensibilité à la corrélation mais dans la direction opposée à celle supposée dans la théorie du regret et de la saillance. L'analyse en grappes identifie une minorité cohérente de sujets sensibles à la corrélation qui influent sur les schémas agrégés, la majorité ne montrant aucune sensibilité à la corrélation. De manière cruciale, l'analyse ne produit pas un type de théorie du regret/saillance. Nous distinguons la sensibilité à la corrélation résultant de comparaisons délibérées au sein de l'état des comparaisons incidentelles des paiements dus à la formulation des problèmes de décision. Les deux produisent une sensibilité à la corrélation, les comparaisons délibérées exerçant une influence quelque peu plus grande.

Dans le deuxième chapitre, également en collaboration avec Jiakun Zheng, nous réexaminons des expériences récentes qui semblent montrer des preuves de sensibilité à la corrélation comme supposé dans la théorie de la saillance. Cependant, ces études échouent à contrôler les effets de division d'événement (ESE). Nous cherchons à démêler le rôle de la corrélation et de la division d'événement dans deux contextes : 1) le paradoxe Allais des conséquences communes tel qu'étudié par Bordalo et al. (2012), Frydman and Mormann (2018), Bruhin et al. (2022) 2) les choix entre paires de Mao tels qu'étudiés par Dertwinkel-Kalt and Köster (2019). Dans les deux contextes, nous trouvons des preuves suggérant que les résultats récents soutenant les effets de corrélation sont largement motivés par l'ESE. Une fois le contrôle de l'ESE effectué, nous ne trouvons aucune preuve cohérente d'effets de corrélation. Nous concluons que nos résultats remettent ainsi en question la validité de la théorie de la saillance dans la description du comportement risqué.

Dans le troisième chapitre, je quitte le domaine de la prise de décision individuelle et considère un cadre de prise de décision déléguée. Dans une expérience guidée par la théorie, j'étudie comment le biais de résultat, une tendance des mandants à récompenser et punir les agents économiques comme s'ils pouvaient avoir anticipé un état aléatoire du monde, façonne les incitations et les choix des agents. Les agents choisissent entre deux loteries au nom de leur mandant. Une loterie est stochastiquement dominante du premier ordre, mais la loterie dominée a plus de chances de

produire un payoff plus élevé état par état. Bien que les mandants observent parfaitement le choix de l'agent, ils ont tendance à récompenser les agents s'ils choisissent la loterie qui réalise un payoff plus élevé. En conséquence, ils incitent les agents à choisir la loterie dominée. Bien que la plupart des agents anticipent ces incitations, seuls les agents stratégiquement sophistiqués ont tendance à choisir la loterie dominée lorsqu'ils pensent avoir une incitation à le faire. L'estimation structurelle suggère que les mandants sont soit totalement biaisés par le résultat, soit totalement impartiaux, les mandants moins cognitivement sophistiqués présentant plus de biais de résultat. Les résultats impliquent que le biais de résultat pourrait être le plus pertinent dans des context.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In my Ph.D. thesis, I study how people make decisions under risk and how they react to and interpret random outcomes. Risk is an important aspect of economic decisions. This is true for the more obvious examples like investments, the stock market, and insurance, which are readily associated with risk. However, many other important decisions people make relating to fields such as education, health, or occupational choice can often be conceptualized as choices between different risky options. Decision-making under risk has captivated economists and their predecessors for at least since the 18th century when Daniel Bernoulli stumbled upon the now famous St. Petersburg paradox[1], which led him to propose that risk prospects should be evaluated by calculating their expected utility (Bernoulli, 1954). Fast forward to 1947, Johann von Neumann and Oscar Morgenstern showed that satisfying a set of four intuitively appealing axioms of rational choice is equivalent to evaluating prospects according to expected utility theory (EUT) (Neumann and Morgenstern, 1953). EUT has been widely accepted as the normative theory of decision-making under risk. It is difficult to exaggerate just how important EUT has become for modern economics - not just as a theory to understand decision-making under risk - but as a fundamental building block in many economic models.

For some time, EUT has also been accepted as an adequate descriptive model of behavior by many, not just a normative one (Bleichrodt and Wakker, 2015). However, at least since Allais (1953), evidence of violations of EUT started to accumulate (Starmer, 2000). Importantly, many of these deviations were systematic in nature, meaning that they seemed to follow certain patterns and led many economists to grow dissatisfied with EUT. Thus, the development of alternatives to EUT started. Given how central EUT is to the field of economics, this venture has attracted considerable attention. What is at stake is not just a better understanding of decision-making under risk as an end in itself but. If one changes the fundamental building block that EUT is to many economic models, this can drastically change the conclusions of said models and improve our understanding of economic behavior in a range of settings. Examples range from asset pricing (Gollier and Salanié, 2006; Bordalo et al., 2012) to auctions (Engelbrecht-Wiggans and Katok,

[1] Consider the following gamble. I toss a coin repeatedly until it comes up tails. I promise you 1$ if the coin comes up heads on the first coin flip, an additional 2$ for the second coin flip, 4$ for the third coin flip, and so on. The paradox is that the expected value of this gamble is infinity, although "any reasonable man would sell his chance, with great pleasure, for twenty [$]" (Bernoulli, 1954, p.31).

2008; Filiz-Ozbay and Ozbay, 2014), to contract theory (Kőszegi, 2014).

In the context of this grand venture of developing a more accurate understanding of decision-making under risk, my Ph.D. thesis can be understood as investigating one specific property of decision-making under risk in which choices can systematically deviate from EUT. This property is correlation-sensitivity.

Choices are correlation-sensitive if they cannot be understood only by reference to the marginal distribution of different prospects, but one has to consider the joint distribution of payoffs. A simple example will go a long way in understanding correlation sensitivity. A decision-maker has to choose between lottery A and B as depicted in Table 1.1. Both lottery A and B yield a payoff of 3$, 2$, or 1$ with equal probability. If, as in EUT, one considers only the marginal distribution of the two options, they are identical and distinguishable only by their respective labels. However, if one considers the joint distribution of payoffs, the options become distinguishable, even if one were to remove their labels. Lottery A yields 3$ whenever B yields 1$. Whenever A yields 2$, B pays 3 $. And whenever A pays 1$, B pays 2$. The work that makes up this thesis is concerned with the following question. Is the joint distribution of payoffs an important feature that economists should consider when studying decision-making under risk? If yes, how does it impact decision-making?

	$s = 1$ (1/3)	$s = 2$ (1/3)	$s = 3$ (1/3)
A	3 $	2 $	1 $
B	1 $	3 $	2 $

Table 1.1 An example of the same-marginal lottery task.

To sharpen the concept of correlation sensitivity it makes sense to delineate it from well-understood reasons for why the joint distribution of different payoff streams matters even when choices are not correlation-sensitive. Diversification in financial markets is a prime example. If an investor has to decide how to allocate money between two assets, how payoffs are correlated matters even if the investor does not care about the joint distribution per se. This is because the joint distribution determines the menu of marginal distributions the investor can obtain by combining different assets in her portfolio. Once presented with the menu of marginal distributions the investor can effectively choose from, the joint distribution of payoffs holds sway over choices only if the investor is correlation-sensitive.

The major avenue to understanding correlation sensitivity is through decision theory. There are two major theories that imply correlation-sensitive preferences, regret theory (Loomes and Sugden, 1982; Bell, 1982; Loomes and Sugden, 1987) and salience theory (Bordalo et al., 2012). In both theories, correlation sensitivity is introduced because decision-makers compare payoffs state-by-state, although for different reasons. In regret theory, once uncertainty is resolved, the decision-maker compares her payoff to the payoff she could have obtained, had she acted differently. If she could have received a higher payoff, she experiences regret, a negative aversive emotion. When choosing between different options, a regret-averse decision-maker anticipates this regret and incorporates it into her utility maximization. In salience theory, the attention of decision-makers is drawn to salient states, that is, states that capture the decision-maker's attention. Importantly,

how salient a state appears to the decision-maker is determined by comparing the payoffs within this state.

Within the realm of theories of decision-making under risk, correlation-sensitive theories form a distinct class of preferences. Most other theories that have been proposed as alternatives to Expected Utility Theory (EUT) such as cumulative prospect theory (Tversky and Kahneman, 1992) or disappointment aversion Gul (1991) retain transitivity. Correlation-sensitive preferences are fundamentally intransitive. Indeed, as we show in chapter 1, in the general framework of correlation-sensitive preferences proposed by Lanzani (2022), preferences are correlation-sensitive if and only if they violate transitivity. This result obtains under relatively weak assumptions - relaxed notions of a complete ordering of alternatives and independence - and highlights the tight connection between correlation sensitivity and transitivity. It is difficult to exaggerate how fundamental transitivity is to standard economic analysis. In a review of non-expected utility theories, (Starmer, 2000, p.362) calls it "a central principle of most economic theory", and (Bleichrodt and Wakker, 2015, p.493) describes it as "one of the most basic assumptions of economic optimisations". Whether preferences are correlation-sensitive, therefore, touches upon very fundamental aspects of decision-making under risk.

In my Ph.D. thesis, I employ lab experiments to study behavior. Lab experiments are a crucial vehicle for theory testing. They enable the researcher to confront participants with choice tasks that are specifically designed to illuminate specific aspects of behavior and to do so in a tightly controlled environment. Observations of choices individuals make in the natural world usually are not of the nature that would allow a researcher to test specific aspects of a theory satisfactorily. By way of example, consider again the choice tasks displayed in Table 1.1. As I will explain below, the choice task can be used to diagnose an individual's correlation sensitivity. Where in the natural world would you find a choice situation like this?[2] It is very telling that the St. Petersburg paradox that led to the development of EUT was observed in the context of an artificially constructed gamble, much like the choice tasks I employ in the experiments that comprise this thesis.

I study whether and how decision-making under risk is correlation-sensitive in two different settings. The first two chapters consider individual decision-making in a setting that is routinely used in the study of risk preferences. Individuals make decisions for themselves. In essence, they are confronted with choices similar to that displayed in Table 1.1, and are asked to choose between the available options. The link to decision-theoretical concepts like correlation sensitivity is rather obvious. In Chapter 3, I study correlation sensitivity in decision-making for others. As I will detail below, in this setting, correlation sensitivity might arise, although no one involved in the interaction cares about correlation sensitivity per se. Yet, one person might act in a way that makes correlation-sensitive decision-making optimal for another person. While decision-theoretic

[2] A criticism of lab experiments is that because one often considers situations that are unlikely to occur in the real world, behavior in the lab is unlikely to reveal anything meaningful about behavior outside the lab. I believe this is rather like saying that a VO2 max test carried out on an assault bike is unlikely to reveal anything about a person's fitness outside of the gym or saying that measuring how well a person can solve a number of Raven's matrices doesn't reveal anything about their ability to reason abstractly about problems they might face in the natural world. That being said, there is a literature on the external validity of lab experiments (see, for instance, the papers reviewed in Camerer (2011)), but this is not the concern of this thesis. For discussion, see Levitt and List (2007); Camerer (2011); Kessler and Vesterlund (2015).

preference models are still useful for thinking about behavior in this setting, chapter 3 ventures out of the realm typically studied in decision-theoretic research. In the following, I summary the main insights from each of the three chapters of the thesis.

Chapter 1: Chapter 1 draws on Loewenfeld and Zheng (2023). We study correlation-sensitive preferences in the general framework of Lanzani (2022). This framework nests regret and salience theory as special cases. Building on Lanzani's work, we show that three distinct types of correlation sensitivity exist. These preference types correspond one-to-one to the three preference relations over the options presented in Table 1.1 - strict preference for lottery A, strict preference for lottery B, and indifference. This result not only simplifies the different possible preference types conceptually but also implies that simple binary choice tasks can be used to classify experimental subjects according to their type of correlation sensitivity. We leverage this insight in a series of experimental studies in which we provide evidence on the prevalence of the different types of correlation sensitivity, as well as their psychological foundations.

The three preference types are best explained with reference to Table 1.1. As pointed out, both lotteries share the same marginal distribution. In choosing between them, a correlation-sensitive decision-maker faces the following trade-off. Lottery A does better by 2\$ in one state, but Lottery B does better by 1\$ in two states. The decision-maker has to aggregate these within-state payoff differences somehow. The first option is to aggregate them linearly, which we refer to as constant sensitivity to payoff differences (CSPD). In this case, the decision maker is indifferent between lottery A and B. The second alternative is that the decision-maker aggregates payoff differences convexly. We refer to this case as increasing sensitivity to payoff differences (ISPD). In this case, the large payoff difference of 2\$ holds more sway than the two smaller differences of 1\$. The decision maker thus has a strict preference for lottery A. Finally, if the decision-maker aggregates payoff differences concavely, we say that she displays decreasing sensitivity to payoff differences (DSPD). In this case, she strictly prefers lottery B.

Regret and salience theory, the two major correlation-sensitive preference models, impose the same preference type, namely ISPD. Our results highlight that ISPD is the crucial property that distinguishes both theories from EUT. In other words, if one were to impose CSPD on regret and salience theory, they would collapse to EUT.[3] Although regret and salience theory both impose ISPD, this is caused by distinct psychological mechanisms. In regret theory, it is assumed that decision-makers are disproportionately averse to large regret. Thus, a regret-averse decision-maker confronted with the choice in table 1.1 is more concerned with avoiding a regret of 2\$ that occurs in one state than avoiding a regret of 1\$ that occurs in two states. In salience theory, the assumption that implies ISPD is that states with large payoff differences are particularly salient to the decision-maker. A decision-maker described by salience theory would, therefore, overweight the first state and underweight the second and third state relative to their objective probability of occurring. This leads the decision-maker to choose lottery A.

Using tasks similar to that depicted in Table 1.1, we run a series of experiments. To probe

[3] The view that regret and salience theory are, in theoretical terms, rather similar is supported by Herweg and Müller (2021), who show that original regret theory (Loomes and Sugden, 1982) is a special case of salience theory (Bordalo et al., 2012), which in turn is a special case of general regret theory (Loomes and Sugden, 1987).

the strength of correlation sensitivity, we also include tasks for which we add a constant payoff premium to either lottery A or B. In addition, we incorporate three between-subject treatments that probe the psychological mechanisms behind correlation sensitivity. These treatments build on the distinct reasons why decision-makers compare payoffs state by state. In regret theory, decision-makers compare payoffs within-state because jointly realized payoffs matter for the decision-maker's utility. Payoff comparisons are thus *deliberate* and should lead to correlation sensitivity only if payoffs realize jointly. In salience theory, decision-makers compare payoffs within the state simply because they have a habit of doing so or because the choice problem is put to them in a way that encourages such comparisons. There is nothing about joint payoff realizations that matters intrinsically. Payoff comparisons might thus be best described as *incidental* and could induce correlation sensitivity even if the compared payoffs do not realize jointly. In order to distinguish between correlation sensitivity due to deliberate and incidental payoff comparisons, our treatments manipulate whether payoffs that are displayed in pairs realize jointly or not.

At the aggregate level, we find evidence for moderate DSPD. At the individual level, we find that most participants are correlation-insensitive, which amounts to EUT in our setting. A minority of around 17% displays consistent DSPD. Importantly, our analysis suggests that virtually none of our participants display ISPD. We further find that both deliberate and incidental payoff comparisons contribute to the observed correlation sensitivity, although deliberate payoff comparisons play a somewhat more important role.

Our results strongly reject ISPD as the prevalent property governing correlation sensitivity. As such, they reject the fundamental behavioral assumption made in both regret and salience theory. It is ISPD that enables the theories to rationalize important deviations from EUT. Much of the literature on non-expected utility models is motivated by exploring alternatives to EUT that are more descriptive and can rationalize systematic deviations from EUT. A major implication of our results for the broader decision-theoretical literature on decision-making under risk is thus that neither Regret nor Salience Theory is a viable alternative to EUT. This is not to say that correlation sensitivity is not an interesting or important phenomenon. However, our results suggest that it is not the driver behind the phenomena that decision theorists are typically interested in rationalizing.

Chapter 2: Chapter 2 might be best understood as a comment on recent research motivated by salience theory. A number of studies tested for correlation sensitivity and reported evidence in favor of ISPD. However, a closer inspection of the experimental design revealed that the reported effects might actually be caused not by correlation sensitivity, but by changes in the choice display that were introduced simultaneously with changes in the joint distribution of payoffs. The confounding effects have first been documented by Starmer and Sugden (1993) in the context of testing regret theory and have been labeled event-splitting effects (ESE). In chapter 2, which builds on Loewenfeld and Zheng (2024), we follow the example of Starmer and Sugden (1993) and disentangle between ESE and correlation sensitivity in the context of choice tasks employed in recent studies. We find that, once ESE are controlled for, no evidence of ISPD remains.

What are ESE? In the recent experiments we consider (Bordalo et al., 2012; Frydman and

Mormann, 2018; Dertwinkel-Kalt and Köster, 2019; Bruhin et al., 2022), subjects had to choose between two lotteries. The treatment variation of interest is a change in the correlation structure while holding the marginal distribution of the lotteries constant. Unfortunately, simultaneously with changing the correlation structure, the number of states displayed to subjects changed. This happens because the number of states changes with the correlation structure. While this change in the choice display should not influence behavior according to salience or regret theory Starmer and Sugden (1993) showed that it can lead to effects similar to what would be expected under ISPD.

To disentangle ESE from correlation sensitivity we build on the design of Starmer and Sugden (1993). We employ two between-subject treatments. In the replication treatment, ESE and changes in the correlation structure are introduced simultaneously. In the control treatment, we hold the number and the probability of states displayed to subjects constant while manipulating the correlation structure. We consider these experimental manipulations in the context of tasks that were employed in the relevant studies on salience theory. These include the Allais paradox, as well as Mao pairs.[4]

Our findings suggest that the evidence for ISDP reported in recent papers on salience theory is most likely driven by ESE and not by correlation-sensitive preferences. In the control treatment, we find no systematic evidence for ISPD. With respect to the Allais paradox, we observe choice patterns similar to those reported by (Bordalo et al., 2012; Frydman and Mormann, 2018; Bruhin et al., 2022) in the replication treatment. However, we fail to replicate the results of Dertwinkel-Kalt and Köster (2019) when considering Mao pairs. It is thus reassuring that a replication of our experiment by Dertwinkel-Kalt and Köster (2021) also finds no evidence for ISPD when controlling for ESE. Another, study, independently carried out by Ostermair (2022), reports findings that suggest moderate DSPD at the aggregate level. Thus, the overall evidence does not support ISPD as the predominant property governing correlation sensitivity.

Chapter 3: In chapter 3, I take the concept of correlation sensitivity out of the realm of individual decision-making and show it can be fruitfully applied to study delegated decision-making. A large empirical literature suggests that decision-makers, such as politicians, CEOs, or financial professionals, are often rewarded or punished for random outcomes. Crucially, this seems to be the case even when outcomes are not informative about the decision-maker or the actions she has taken. Evidence from tightly controlled lab studies points in a similar direction. Principals show a strong tendency to reward agents at least partially for random outcomes, even when the agent's choice is known to principals and outcomes are literally determined by the flip of a coin (Gurdal et al., 2013; de Oliveira et al., 2017; Brownback and Kuhn, 2019; Aimone and Pan, 2020; König-Kersting et al., 2021). The tendency to reward agents based on random, uninformative outcomes is often referred to as Outcome Bias.

The main question I seek to address in chapter 3 is how the Outcome Bias of Principals might influence the monetary incentives, and, ultimately, their choices. Outcome Bias is often regarded

[4] Mao pairs are pairs of lotteries that have the same expected value and the same variance. However, the skewness of one lottery equals the negative of the skewness of the other lottery.

as undesirable and potentially harmful. It violates a central result in the theoretical principal-agent literature that principals seeking to incentivize agents to make good choices should never reward or punish them for random, uninformative outcomes (Holmström, 1979). However, despite the strong evidence that Outcome Bias is prevalent in a range of settings, including settings where bad decisions could lead to large welfare losses, little is known about how agents behave if they are being rewarded and punished in an outcome-biased way.

How does correlation sensitivity enter the picture? The punchline of Chapter 3 is that Outcome Bias can induce principals to incentivize agents as if their preferences were correlation-sensitive, even if they are not correlation-sensitive. This can induce principals to de facto incentivize agents to choose their least preferred option from a pair of lotteries. If agents anticipate the principal's Outcome Bias, they might thus be tempted to choose an option for the principal that harms her welfare.

I consider the following stylized principal-agent interaction. The agent ("he") has to choose between two lotteries on behalf of the principal ("she"). The lotteries are similar to those displayed in Table 1.1. However, I add a payoff premium to all payoffs of either lottery A or B. This ensures that there is always one lottery that is clearly better in a normative sense. The "good" lottery dominates the "bad" lottery in the sense of first-order stochastic dominance. The returns of the agent's choice accrue exclusively to the principal. The principal observes the choice of the agent and she also observes the realized state of the world. Then, she has to decide whether to award a fixed bonus to the agent or not. The setting is chosen to capture crucial aspects of situations where agents face ex-post evaluations by a principal. The principal can essentially reward the agents based on two criteria, her choice or the realized outcome. The agent does not know how the principal will reward him when he makes his choice, but he has to form expectations. This feature is prominent in virtually all applications of interest and enables me to study how agents form expectations about the principals' outcome bias. Note also that this interaction is free of any of the principal-agent problems typically considered in the literature. The first-best could be obtained quite trivially, which puts the focus on frictions arising due to outcome bias.

I start from the observation made by Gurdal et al. (2013) that bonus payments tend to be based on a comparison between the outcome the agent actually obtained and the outcome he could have obtained, had he chosen a different option. Another way to think about this is that the principal rewards the agents as if he could have anticipated the realization of a random variable. I propose a simple model of Outcome Bias that captures this notion. The principal is motivated by reciprocity to reward the agent for good choices. However, in evaluating whether the agent made a good choice, she focuses too much on the realized states and neglects non-realized states. In the case of full outcome bias, she simply awards the bonus only if the agent chooses the lottery that yields the higher payoff in the realized state.

This way of rewarding can create perverse incentives. Consider again the choice task in Table 1.1. A fully outcome-biased principal will reward the agent for choosing lottery A if state 1 materialized, but she will reward him for choosing lottery B in states 2 and 3. Overall, this principal is more likely to award the bonus if the agent chooses lottery B - she de facto provides

incentives for the agent to choose lottery B. This is even be true if we add a payoff premium (smaller than 1 \$) to each payoff of lottery A, which renders lottery A first-order stochastic dominant. The principal might now have a strict preference for lottery A, and yet incentivize the agent to choose lottery B.

I show formally (in the appendix to Chapter 3) that the reason why outcome bias can create perverse incentives is that it creates incentives that are correlation-sensitive, even if the principal's preferences are not. Consider a principal whose preferences are described by EUT and an agent who perfectly anticipates the principal's bonus decisions and chooses between two lotteries in a way that maximizes his monetary payoffs. I show that the agent's choices can be represented by a utility function that satisfies the axioms of Lanzani's framework of correlation-sensitive preferences. Crucially, this utility function satisfies EUT if and only if the principal does not display any outcome bias, i.e., if the principal rewards the agent only based on choices. If, however, the principal displays outcome bias, i.e. rewards the agent based on an ex-post comparison of the obtained and the forgone outcome, the utility function is correlation-sensitive.

In an experiment I ran at the Toulouse School of Economics, I find that exactly this can happen. When asked to choose for themselves, participants in the role of the principal choose the dominant lottery at a very high frequency, regardless of which lotteries yield a higher payoff in more states. However, principals show a strong tendency to reward the agent if he chooses the lottery that yields a higher payoff. As a result, many principals are more likely to reward the agent for choosing the bad lottery if it yields a higher payoff in two out of three states - they de facto incentivize agents to choose their least preferred lottery.

Based on stated beliefs, most agents anticipate the principals' outcome bias well. However, most agents do not best respond to their own beliefs and choose the good lottery anyhow. A closer inspection of the data reveals an intriguing pattern. Agents who are - based on survey responses - arguably more strategically sophisticated do tend to choose the bad lottery more often when they believe in having incentives to do so, but only after they have been nudged to form beliefs through belief elicitation. This finding suggests that a certain level of strategic sophistication on the part of the agents might be necessary for anticipated outcome bias to translate into actions that are detrimental to the principal.

I also estimate the proposed model of outcome bias structurally. This yields three additional insights. First, using both individual-level estimation and finite mixture models, I find that most principals are either fully outcome-biased or fully unbiased. Second, I find evidence that high levels of outcome bias are linked to low levels of cognitive reflection. This suggests that outcome bias might work through a cognitive channel, as the model assumes. It also suggests that outcome bias might be most prevalent in settings, where principals are relatively unsophisticated. Third, when estimating the model on agents' stated beliefs, I find that agents tend to somewhat overestimate how outcome biased principals are.

The results regarding individual heterogeneity in principals' propensity to display outcome bias help paint a very complete picture. Overall, my results suggest that outcome bias is most likely to induce agents to take actions detrimental to the principal in situations where agents are sufficiently

sophisticated but principals lack such sophistication. I would argue that this description applies to many if not most, settings of delegated decision-making. After all, the reason why we delegate decisions to others is often because they possess expertise that we lack. Sophisticated individuals are likely to select themselves into important positions in which they act as agents, may that be in politics, corporate governance, or investment management. Principals, such as voters or ordinary citizens seeking a mutual fund to invest for retirement, might often lack sophistication, especially in the agent's domain of expertise.

Chapter 2

Uncovering Correlation Sensitivity in Decision Making under Risk

Moritz Loewenfeld[1] Jiakun Zheng[2]

Abstract

Allowing risk preferences to be sensitive to the correlation between lottery outcomes can explain classical deviations from expected utility theory as well as phenomena in various real-world settings. However, evidence on correlation sensitivity is limited and mixed. We show that correlation-sensitive preferences in the framework of Lanzani (2022) can be classified into three categories. We propose a choice task to classify experimental subjects accordingly. In multiple experiments, we find that aggregate choices display correlation sensitivity but in the opposite direction as assumed in regret and salience theory. Clustering analysis identifies a consistent correlation sensitive minority driving aggregate patterns, with the majority showing no correlation sensitivity. Crucially, the analysis does not produce a regret/salience theory type. We disentangle correlation sensitivity arising from deliberate within-state comparisons from incidental payoff comparisons due to the framing of decision problems. Both produce correlation sensitivity, with deliberate comparisons exerting a somewhat greater influence.

Keywords: Choice under Risk; Correlation Effects; Experiment; Regret Theory; Salience Theory

JEL Classifications: C91; D81; D91

[1] Toulouse School of Economics, University Toulouse Capitole, 1 Esplanade de l'Université, Toulouse 31080 Cedex 06, France. Email(✉): moritz.loewenfeld@tse-fr.eu.
[2] Aix Marseille Univ, CNRS, AMSE, Centrale Marseille, Marseille, France. Email (✉): jiakun.zheng@outlook.com.

2.1 Introduction

In an influential class of risk preferences including regret theory and salience theory (Bell, 1982; Loomes and Sugden, 1982; Bordalo et al., 2012), the correlation between different risky prospects can significantly impact choices. Incorporating correlation sensitivity in preferences can help resolve several deviations from expected utility theory (EUT), including the Allais paradox (Allais, 1953), preference reversals, or simultaneous gambling and insurance. Moreover, it provides an explanation for skewness preferences (Dertwinkel-Kalt and Köster, 2019) and asset price puzzles (Bordalo et al., 2013).

However, the key behavioral implications of correlation sensitivity remain somewhat elusive, and existing experimental evidence on the prevalence and nature of correlation sensitivity is inconclusive. In this paper, we consider the general class of correlation-sensitive preferences axiomatized by Lanzani (2022), which nests regret (Bell, 1982; Loomes and Sugden, 1982, 1987) and salience theory (Bordalo et al., 2013). We show that there exist three types of correlation sensitivity. We then propose a simple experimental task that allows us to classify experimental subjects according to their type of correlation sensitivity. Finally, we provide experimental evidence on the prevalence and nature of correlation sensitivity, and probe its psychological foundations. Our experimental task, the same marginal lotteries (SML) task illustrated in Table 2.1, captures the key features of correlation sensitivity. The marginal distribution of both lotteries is described by three distinct payoffs: h, m, and l, each of which occurs with equal probability. Whenever the row lottery yields the payoff h, m, or l, the column lottery yields, respectively, the payoff l, h, or m. Note that both lotteries share the same marginal distribution and are distinguishable only by the way payoffs are correlated. The row lottery yields a much higher outcome than the column lottery with probability $1/3$ (h vs l), but a somewhat lower outcome with probability $2/3$ (m vs h and l vs m). Evidently, a decision maker whose preferences are fully characterized by the marginal distribution of payoffs is indifferent between the two lotteries. A correlation-sensitive decision maker, however, has a strict preference for either the row or the column lottery.

Intuitively, the decision maker has to aggregate the differences between joint payoff realizations to form a preference. If she aggregates payoff differences linearly, we say that she satisfies constant sensitivity to payoff differences (CSPD). In this case, she is indifferent between the row and the column lottery. If the decision maker is increasingly sensitive to payoff differences (ISPD), large payoff differences have an outsized effect on preferences. The decision maker factors the high-contrast realization (h, l) more heavily than the two smaller differences of the realizations (m, h) and (l, m) combined. As a result, the decision maker has a strict preference for the row lottery. If the decision maker is decreasingly sensitive to payoff differences (DSPD), the reverse applies.

π^{sml}	$h(1/3)$	$m(1/3)$	$l(1/3)$
$h(1/3)$	0	1/3	0
$m(1/3)$	0	0	1/3
$l(1/3)$	1/3	0	0

Table 2.1 The same marginal lotteries (SML) task

Heuristically speaking, a decision maker characterized by ISPD chooses the lottery that yields the higher outcome for the payoff realization with the highest contrast, whereas a decision maker characterized by DSPD chooses the lottery that yields a higher payoff most of the time.

In the framework of Lanzani (2022), we show that preferences over the lotteries of the SML task allow us to classify people into the three mutually exclusive categories of preferences characterized by CSPD, ISPD, and DSPD. Lanzani provides a representation theorem for the class of correlation-sensitive preferences. He further shows that if transitivity is imposed, correlation-sensitive preferences collapse to EUT and become insensitive to the correlation structure. We slightly extend Lanzani's result to show that transitivity is equivalent to a simple and intuitive condition on the correlation-sensitive utility function. This condition can be understood as imposing CSPD, which implies that preferences are not correlation-sensitive, but are fully described by a relation over marginal distributions of payoffs. It follows that in Lanzani's framework, transitivity can be violated in two ways, and thus correlation sensitivity can take two directional forms: either ISPD or DSPD. We show that a decision maker satisfies CSPD if and only if she is indifferent between the row and the column lottery of the SML task. The decision maker strictly prefers the row lottery (the column lottery) if and only if she satisfies ISPD (DSPD). Therefore, we occasionally refer to the row as the ISPD lottery and the column as the DSPD lottery.

As Lanzani's framework nests regret and salience theory, our result shows that correlation sensitivity is the key behavioral property that distinguishes regret and salience theory from EUT. In other words, if one were to impose CSPD on regret and salience theory, the theories would reduce to EUT. Both theories impose ISPD, though they postulate different psychological mechanisms. In salience theory, joint realizations receive different decision weights according to their salience (Bordalo et al., 2012). The assumption that implies ISPD is that joint realizations with large payoff differences are the most salient and thus receive disproportionate decision weight. In regret theory, the decision maker's utility depends on a comparison of jointly realized payoffs. If the decision maker realizes that they could have received a higher payoff had they chosen differently, she experiences regret (Loomes and Sugden, 1982, 1987). The crucial assumption that yields ISPD is that the decision maker is increasingly sensitive to increments in regret.

The SML task serves as the basis for our experimental tests of correlation sensitivity. Since correlation sensitivity implies strict preferences for either the row or the column lottery, failure to detect correlation sensitivity thus provides strong evidence of correlation insensitivity. Intuitively, since the marginal distribution of both lotteries is the same, even correlation sensitivity of otherwise negligible importance should be apparent. On the flip side, if choices display correlation sensitivity, this might be caused by a weak preference that might not be evident in any other decision problem. To probe the strength of potential correlation sensitivity, we add a payoff premium to all payoffs of either the ISPD or the DSPD lottery, rendering one of the lotteries First-Order Stochastic Dominant (FOSD). We refer to these decision problems as FOSD tasks. A decision maker with CSPD has a strict preference for the lottery which is FOSD. Correlation-sensitive decision makers may violate FOSD in one direction, which comes at the cost of forgoing higher payoffs.

We further implement three between-subject treatments to investigate the psychological foun-

dations of correlation sensitivity. ISPD is the key theoretical property in both regret and salience theory, but these theories rely on distinct psychological mechanisms. In regret theory (Bell, 1982; Loomes and Sugden, 1982), payoff comparisons are deliberate because the decision maker's true utility depends on a comparison of jointly realized payoffs. In salience theory (Bordalo et al., 2012), on the other hand, payoff comparisons impact the decision maker's perception of the choice at hand. As there is nothing about joint payoff realizations that the decision maker values intrinsically, payoff comparisons might be best described as incidental. Different authors have suggested that the framing of decision problems, rather than the joint realization of payoffs, might be the relevant criterion to determine the unit of payoff comparisons in salience theory (Dertwinkel-Kalt and Köster, 2015; Leland et al., 2019).

Our between-subject treatments build on the distinction between deliberate comparisons of jointly realized payoffs and incidental payoff comparisons due to the framing of the decision problem. In all treatments, the joint distribution of lotteries is described by reference to states of nature, which are determined by the turn of a wheel of fortune, and the lotteries yield different outcomes depending on the realized state of nature. We present the choice problems in tabular form. In the baseline treatment, each column describes the realizations of each lottery in a given state of the world. In this treatment, correlation sensitivity can arise from deliberate state-by-state comparisons of payoffs or incidental column-by-column comparisons of payoffs. Our remaining two treatments allow us to disentangle the two channels. In the column-effects treatment, payoffs are perfectly correlated across states, which means correlation sensitivity can only arise from column-by-column comparisons of payoffs. On the other hand, in the state-effects treatment, the payoffs displayed in each column are the same for both lotteries, while the joint distribution of payoffs remains consistent with the baseline treatment. Consequently, correlation effects observed in this treatment can only be attributed to deliberate state-by-state comparisons of payoffs across columns.

We further supplement the SML task in the baseline treatment with a within-subject treatment that varies whether subjects receive immediate outcome feedback after a given choice. We include this treatment to accommodate for the contingency that regret aversion is a prominent feature in decision making only if individuals anticipate immediate outcome feedback (Bell, 1983; Zeelenberg et al., 1996; Zeelenberg, 1999).

In one lab and two online experiments, comprising a total of 919 participants, we collected more than 18,000 observations. In the baseline treatment, the aggregate choices provide evidence for small but consistent effects that imply DSPD. We also document a precise null effect of immediate outcome feedback, suggesting that it does not interact with correlation sensitivity in a significant way. Using latent class analysis to identify categories of choice patterns, we find that most of our participants exhibit behavior consistent with CSPD. However, a minority of around 17% of the participants consistently exhibit behavior that satisfies DSPD, even if it violates FOSD. Importantly, the latent class analysis does not produce a corresponding category of participants whose behavior is characterized by ISPD. Therefore, both our aggregate and individual-level results strongly reject ISPD as the prevalent property governing preferences.

We find further evidence of DSPD in both the column-effects and the state-effects treatments.

However, compared to the baseline treatment, the aggregate effect size is somewhat reduced, particularly in the column-effects treatment. Furthermore, when analyzing individual responses, we find that only 9% of participants in the column-effects treatment are assigned to the latent class characterized by strong DSPD. In contrast, the corresponding fraction in the state-effects treatment is 22%. This indicates that deliberate comparisons of jointly realized payoffs may play a somewhat more significant role in driving correlation sensitivity compared to incidental column-by-column comparisons of payoffs.

2.1.1 Related literature

We contribute to three strands of literature. First and foremost, we contribute to the experimental literature on correlation-sensitive preferences. Although models of correlation-sensitive preferences have drawn interest at least since the proposal of regret theory (Bell, 1982; Loomes and Sugden, 1982) and have recently seen a revival in salience theory (Bordalo et al., 2012), the experimental evidence for correlation sensitivity is inconclusive.

Three approaches to examining correlation sensitivity can be distinguished. The first approach builds on manipulations of the joint distribution of payoffs. A number of papers motivated by testing regret theory (Loomes and Sugden, 1987; Loomes, 1988a,b) and salience theory (Bordalo et al., 2012; Frydman and Mormann, 2018; Dertwinkel-Kalt and Köster, 2019; Bruhin et al., 2022) used this approach and reported choice patterns that purportedly provided evidence for ISPD. However, Starmer and Sugden (1993) showed that the initial results attributed to correlation sensitivity were most likely caused by so called event-splitting effects, which are unintended changes in the choice display.[3] Once controlling for these simultaneous changes in the choice display, Starmer and Sugden (1993) found that the evidence for correlation sensitivity was considerably weakened and lost statistical significance. Humphrey (1995), and more recently Ostermair (2021) and Loewenfeld and Zheng (2024) also failed to find evidence for correlation sensitivity once changes in the choice display were controlled for. However the design of these studies does not allow to conclude that preferences satisfy CSPD. We will elaborate on this point in section 2.3.2. Therefore, their findings may be inconclusive. This view is echoed by (Starmer and Sugden, 1993, p.253), who find no statistically significant evidence for correlation sensitivity but argue that their "*data display a clear tendency towards a* [correlation] *effect, and it may be that such effects would be more apparent in other problem settings.*"

A second approach seeks to measure the correlation-sensitivity of preferences in a non-parametric way using the trade-off method (Wakker and Deneffe, 1996). Adopting this approach, both Bleichrodt et al. (2010) and Baillon et al. (2015) reported evidence suggesting the majority of their participants satisfied ISPD. Unfortunately, these results are subject to severe limitations. First, the trade-off method involves a dynamically generated series of choices. The dynamic nature of

[3] In these studies, subjects were confronted with a choice between two different lotteries under two different joint distributions. Correlation sensitivity implies that subjects might shift their choices in response to the change in the joint distribution. However, simultaneous to changing the correlation structure, the number of states displayed to subjects was changed as well, in a way that has been shown to produce behavioral patterns similar to those implied by ISPD (Starmer and Sugden, 1993; Loewenfeld and Zheng, 2024).

the method creates a critical issue, namely a lack of incentive compatibility, which could produce behavior equivalent to ISPD.[4] One might argue that subjects are unlikely to understand the lack of incentive compatibility, but ultimately, this remains a question of faith. Even if one believes incentive compatibility is not an issue, there are reasons to doubt that the trade-off method produces reliable measures of correlation sensitivity. First, Andersson et al. (2023) failed to replicate the findings of Bleichrodt et al. (2010) and Baillon et al. (2015) with M-turk workers, who are supposedly more representative than university students.[5] Second, Baillon et al. (2015) found that, on a different set of choice tasks, subjects displayed behavior largely incompatible with their measure of correlation sensitivity obtained with the trade-off method.

The third approach relies on testing for preference cycles as prescribed by regret theory (Loomes et al., 1991; Baillon et al., 2015). As pointed out above, in Lanzani's framework, decision makers violate transitivity if and only if they are correlation-sensitive.[6] Therefore, tests for preference cycles predicted by ISPD constitute a test of correlation insensitivity. Initial studies seemed to find support for such cycling preferences (Loomes et al., 1989, 1991). However, subsequent studies claimed that the preference cycles previously observed were likely the result of decision noise rather than intransitive preferences (Sopher and Gigliotti, 1993; Regenwetter et al., 2011). Motivated by this development, Baillon et al. (2015) employed a two-step procedure. In the first step, they employed the trade-off method to measure subjects' correlation-sensitive preferences. In the second step, subjects were confronted with choice-triples tailored to their preferences, such that systematic violations of transitivity should be triggered. However, this was not observed, despite simulation exercises suggesting statistical power near 100%. These findings suggests that the trade-off method might not provide reliable estimates of correlation sensitivity.

With the SML task, which was previously used by Leland (1998) to illustrate theoretical properties of regret theory, we advance the investigation of correlation sensitivity by providing a novel diagnostic tool. From an experimental viewpoint, the simplicity of the SML task is its main strength as it avoids methodological shortcomings. It is incentive compatible, and the results can be interpreted in a straightforward way, while undesirable and confounding features of the choice display are naturally avoided (see section 2.3.2 for details). The SML task is capable of detecting correlation sensitivity, even if it were of negligible importance, which is a theoretical advantage. Moreover, as a classification task, it is ideally suited to study correlation sensitivity not only on the aggregate, but also the individual level.

While previous studies have been somewhat inconclusive, these property of the SML task enables us to firmly reject that behavior satisfies ISPD both at the aggregate and the individual level. Furthermore, in a large sample of participants, we document for the first time small but persis-

[4] The trade-off method consists of eliciting a series of values x_j that make the decision maker indifferent between a lottery $(x_j, p; g, 1-p)$ and another lottery $(x_{j-1}, p; G, 1-p)$, with g, G, x_0, and p being chosen by the experimenter. In the first step, g, G, and x_0 are used to elicit x_1. The second step then consists of using g, G, and x_1 to elicit x_2, and so on up to x_5 in Bleichrodt et al. (2010). If subjects anticipate the structure of the method, it provides an incentive to report higher values of x_1, x_2, etc., as would correspond to their actual preferences.

[5] One potential explanation for the difference between Andersson et al. (2023) and the other studies could be attributed to the possibility that online subjects have a lower level of sophistication in comprehending this aspect compared to university students.

[6] The SML task is a "reduced form" test for the cycle $A \succ B \succ C \succ A$ as it cuts out the middle part and directly tests for $A \succ A$.

tent effects implying DSPD at the aggregate level, which is driven by a minority of participants who consistently display behavior satisfying DSPD. These results are in sharp contrast with the predictions of regret and salience theory and existing findings in the literature. We attribute our ability to document DSPD to a mix of high statistical power and our improved experimental task. While we are the first to document moderate DSPD, at least one coherent result emerges from the experimental literature, namely the absence of ISPD. The only experiments producing evidence for ISPD use the trade-off method. As we discussed above, there are good reasons to doubt the validity of these results. It is noteworthy that other previous experiments only suggest the absence of evidence for ISPD at the aggregate level, whereas our classification task provides evidence for the absence of ISPD, even at the individual level.

We also contribute to the experimental literature on correlation sensitivity by investigating its psychological foundations. To the best of our knowledge, we are the first to do so. Our results suggest that both incidental payoff comparisons due to the framing of choice problems, as well as deliberate state-by-state comparisons of payoffs play their part in producing correlation sensitivity, although the latter channel might be somewhat more important.

Finally, we contribute to the broader literature on regret and salience theory. A number of experimental studies have tested the implications of regret (Somasundaram and Diecidue, 2017) and salience theory (Dertwinkel-Kalt et al., 2017; Königsheim et al., 2019; Alós-Ferrer and Ritschel, 2022), but without testing for correlation sensitivity. A large literature has developed regret theory theoretically (Bell, 1982; Loomes and Sugden, 1982; Bell, 1983; Loomes and Sugden, 1987; Quiggin, 1990, 1994; Sarver, 2008; Diecidue and Somasundaram, 2017; Gollier, 2020). Applied theoretical work has explored the implications of regret theory for insurance demand (Braun and Muermann, 2004; Wong, 2012), portfolio choice (Michenaud and Solnik, 2008; Qin, 2015), asset pricing (Gollier and Salanié, 2006), and health prevention (Zheng, 2021). Salience theory (Bordalo et al., 2012) has been shown to provide a potential foundation for skewness preferences (Dertwinkel-Kalt and Köster, 2019) and has been applied to the study of asset pricing, both theoretically (Bordalo et al., 2013) and empirically (Cosemans and Frehen, 2021), as well as the newsvendor problem (Dertwinkel-Kalt and Köster, 2017).

The popularity of these two theories is due to their intuitively plausible explanations for many perplexing choice phenomena observed in various settings, ranging from experimental labs to insurance and asset markets. We contribute to this literature by testing the ISPD property, which is the key theoretical property that sets apart regret and salience theory from other prominent decision-making theories under risk. Our rejection of ISPD at the aggregate and individual level suggests that the mechanisms postulated in regret and salience theory are unlikely to be the cause of deviations from EUT, such as the Allais paradox or simultaneous gambling and insurance, as well as behavioral tendencies like the commonly observed preference for right-skewed risks. One might argue that much of the ability of regret and salience theory to explain these behavioral anomalies is due to its capability to endogenize the probability weighting of cumulative prospect theory (Tversky and Kahneman, 1992). The recent literature on behavioral inattention and Bayesian updating provides promising avenues for understanding these behaviors without violating transitivity

(Gabaix, 2014; Enke and Graeber, 2021).

The remainder of this paper is organized as follows. Section 2.2 provides a formal discussion of correlation-sensitive preferences. We then introduce our experimental task in section 2.3. Section 2.4 details the experimental procedures, and section 2.5 presents the main experimental results. We discuss our results in section 2.6. Section 2.7 concludes.

2.2 Correlation-sensitive preferences

In the framework of Lanzani (2022), choices between two lotteries are described by a nonempty set of payoffs X and a finite measure of the joint probability distribution $\pi \in \Delta(X \times X)$. To avoid technicalities we impose $X \subseteq \mathbb{R}$. Consider Table 2.2. The decision maker decides between the row and the column lottery so as to be paid accordingly to the realized row or column outcome. The payoff pair (x_i, y_j) realizes with a probability of π_{ij}. For a decision maker who is not correlation-sensitive, preferences are fully described by a binary ranking over the marginal distribution of the row lottery $\pi_1 \in \Delta(X)$ (light gray area in Table 2.2) and that of the column lottery $\pi_2 \in \Delta(X)$ (gray area in Table 2.2). Formally, the marginal distribution of the row lottery is given by $\pi_1(x) = \sum_{y \in Y} \pi(x, y) = (p_1, ..., p_m)$ and that of the column lottery by $\pi_2(y) = \sum_{x \in X} \pi(x, y) = (q_1, ..., q_n)$. To allow for correlation sensitivity, Lanzani defines the decision maker's preferences over the joint distribution of outcomes. Binary preferences are modeled as a preference set $\Pi \subseteq \Delta(X \times X)$. The decision maker is said to have a preference for the row lottery for a given joint distribution π if $\pi \in \Pi$. Define $\overline{\pi}$ to be the conjugate distribution of π, that is, $\forall (x, y) \in X \times X$, $\overline{\pi}(x, y) = \pi(y, x)$. Intuitively, the conjugate distribution is the distribution that obtains when switching the row and the column lottery. Then, a decision maker has a preference for the column lottery if $\overline{\pi} \in \Pi$.

The relation $\pi \in \Pi$ is analogous to the familiar weak preference relation $\succeq$, and $\pi, \overline{\pi} \in \Pi$ corresponds to indifference. Note that the classical preference relation $\succeq$ also induces a preference set. However, it is defined over $\Delta(X) \times \Delta(X)$ whereas Π is defined over $\Delta(X \times X)$. Given the notion of weak preferences, a notion of strict preferences, in the language of preference sets, can be introduced. Given a preference set Π, the subset of strict preferences $\hat{\Pi}$ is defined as $\hat{\Pi} = \{\pi \in \Pi : \overline{\pi} \notin \Pi\}$. That is, for $\pi \in \hat{\Pi}$ the decision maker strictly prefers to be paid according to the row rather than the column lottery.

Lanzani (2022) imposes three axioms on the preference set Π that are necessary and sufficient to obtain a representation theorem for correlation-sensitive preferences. The three axioms are completeness, strong independence, and continuity, and are natural analogues to the corresponding

π	$y_1(p_1)$	...	$y_m(p_m)$
$x_1(q_1)$	π_{11}	...	π_{1m}
...	...	...	...
$x_n(q_n)$	π_{n1}	...	π_{nm}

Table 2.2 Binary choices in the tabular form

axioms in the v.N.M axiomatization of EUT.[7] Consider a skew-symmetric function $\phi : X \times X \to \mathbb{R}$, that is, $\phi(x, y) = -\phi(y, x)$, $\forall (x, y) \in \mathbb{R} \times \mathbb{R}$. A preference set Π satisfies Lanzani's three axioms if and only if there exists a skew-symmetric function ϕ such that, for any $\pi \in \Delta(X \times X)$

$$\pi \in \Pi \iff \sum_{x,y} \phi(x, y)\pi(x, y) \geq 0 \tag{2.1}$$

As Lanzani points out, the binary preference relation induced by preferences defined over the marginal distributions of payoffs is a possibly incomplete subset of the relation defined over the joint distribution of payoffs. In other words, preferences defined over the joint distribution are more general than preferences defined over the marginal distribution of payoffs. As such, it might not be surprising that Lanzani's framework can accommodate EUT preferences.[8]

However, the generality of correlation-sensitive preferences comes at a cost of added complexity. Whether this added complexity is necessary, that is whether risk preferences are indeed meaningfully correlation-sensitive, is a crucial question. As it is part of our motivation to provide evidence on this question, we introduce a formal definition of correlation-sensitive preferences, whose counterpart are naturally correlation-insensitive.

Definition of correlation-sensitive preferences. *Let the set of payoffs $X = \mathbb{R}$. A preference relation is correlation-sensitive if and only if $\exists \pi$, $\pi' \in \Delta(X \times X) : \pi_1 = \pi'_2$, $\pi_2 = \pi'_1$, $\pi \in \hat{\Pi}$ and $\pi' \in \hat{\Pi}$.*

Intuitively, preferences are correlation-sensitive if and only if there exists a pair of row and column marginals such that the row lottery is strictly preferred under one joint distribution π, but the column lottery is strictly preferred under a different joint distribution π'. In other words, correlation-sensitive preferences cannot be fully described by a binary relation over marginal distributions.

Lanzani (2022) introduces a fourth axiom, transitivity, which is essentially a translation of the classic transitivity axiom into the language of preference sets. In words, the transitivity states that if a marginal distribution π_1 is preferred to another marginal distribution χ_1 under a joint distribution π, and the marginal distribution χ_1 is preferred to another marginal distribution χ_2 under a joint distribution χ, then the marginal distribution π_1 must be preferred to the marginal distribution χ_2 under a joint distribution ρ.[9]

In his Proposition 1, Lanzani (2022) shows that if transitivity is imposed, the preference relation is fully characterized by an EUT representation. Below, we restate Lanzani's proposition, and slightly extend it by showing that transitivity is equivalent to correlation insensitivity, as well as to

[7] Whereas completeness and archimedean continuity can be seen as a translation of the corresponding axioms in the v.N.M. axiomatization of EUT into the language of preference sets, the strong independence axiom implies considerably less structure than the corresponding standard independence axiom of EUT. The standard independence axiom implies that common consequences, understood as payoffs that are yielded by the marginal distribution of both lotteries, do not impact choice and can be edited out. The strong independence axiom implies that payoffs yielded by both the row and the column lottery can be edited out if they realize jointly.

[8] A preference set admits an EUT representation if there exists $u : X \to \mathbb{R}$ such that $\pi \in \Pi \iff \sum_{x,y}(u(x) - u(y))\pi(x, y) \geq 0$. In other words, $\phi(x, y) = u(x) - u(y)$.

[9] Formally, the transitivity axiom can be stated as follows: $\forall\, \pi, \chi, \rho \in \Delta(X \times X)$, if $\pi_2 = \chi_1$, $\rho_1 = \pi_1$, and $\rho_2 = \chi_2$, then $(\pi \in \Pi, \chi \in \Pi) \Rightarrow \rho \in \Pi$.

a simple and intuitive condition on the function ϕ. This characterization of correlation insensitivity will greatly help in clarifying the directional effect of correlation sensitivity and in deriving our experimental tests of correlation sensitivity.

Proposition 1. *If* Π *admits a correlation-sensitive preference representation as given in the expression* (2.1)*, the following statements are equivalent:*

1. Π *satisfies transitivity.*

2. Π *is fully characterized by EUT (and thus correlation-insensitive).*

3. $\forall h, m, l \in \mathbb{R}$ *such that* $h > m > l$, $\phi(h, l) = \phi(h, m) + \phi(m, l)$

Proof. see Appendix A.1. □

2.2.1 Classification of correlation-sensitive preferences

This restated and extended version of Lanzani's proposition is meaningful for several reasons. First, it highlights that preferences are correlation-sensitive, and transitivity is violated if and only if $\phi(h, l) \neq \phi(h, m) + \phi(m, l)$. This allows to distinguish broadly between three mutually exclusive categories of preferences. The first category is correlation insensitivity. Decision makers falling into this category satisfy Constantly Sensitivity to Payoff differences (CSPD). such decision makers are indifferent between one large payoff difference and two smaller differences that add up to the same size (see Proposition 1). For the other two categories of preferences, correlation sensitivity goes in two opposite directions. If $\phi(h, l) > \phi(h, m) + \phi(m, l)$, decision makers prefer one large payoff difference to two smaller differences that add up to the same size. We will say that decision makers in this category are Increasingly Sensitive to Payoff Differences (ISPD). For the reverse direction, we will say that decision makers are Decreasingly Sensitive to Payoff Differences (DSPD). This forms the third category. We summarize the classification of correlation-sensitive preferences below.

Classification of correlation-sensitive preferences. *Suppose that a decision maker's preference relation admits a correlation-sensitive preference representation as given in the expression* (2.1)*. Then,*
(1) the decision maker is CSPD (or correlation-insensitive) if $\phi(h, l) = \phi(h, m) + \phi(m, l)$*;*
(2) the decision maker is ISPD if $\phi(h, l) > \phi(h, m) + \phi(m, l)$*;*
(3) the decision maker is DSPD if $\phi(h, l) < \phi(h, m) + \phi(m, l)$*.*

The condition on the preference functional connects naturally to the literature on regret and salience theory. In their generalization of original regret theory (Loomes and Sugden, 1982), Loomes and Sugden (1987) impose the condition $\phi(h, l) > \phi(h, m) + \phi(m, l)$, which is usually referred to as regret aversion or convexity in the regret theory literature. As Herweg and Müller (2021) demonstrate, from a mathematical perspective, salience theory is a special case of generalized regret theory. Thus, both salience and and regret regret theory imply ISPD. Proposition 1 highlights the fact that correlation-sensitivity is the defining feature of both salience and regret

theory. When preferences satisfy CSPD, they are correlation-insensitive and collapse to an EU representation.

Finally, it is worth mentioning that in the context of binary choices, the recently proposed attention theory by Chew et al. (2023) incorporates correlation sensitivity. Particularly, when the attention function exhibits skew symmetry, attention theory aligns with regret theory and salience theory. While our experiments primarily focus on characterizing correlation sensitivity within the framework of Lanzani (2022), they also serve as a test for correlation sensitivity induced by attention theory.

2.3 An experimental test of correlation sensitivity

2.3.1 The same marginal lotteries (SML) task

As discussed in the introduction, Table 2.1 presents our main experimental task, the SML task. As the name suggests, subjects choose between two lotteries that have the same marginal distribution. The lotteries can be distinguished only based on the joint distribution π^{sml}. In Table 2.3i and 2.3ii, a payoff premium (i.e., $\epsilon > 0$) is added to all possible payoffs of either the row or the column lottery, which makes the corresponding lottery first-order stochastic dominant. We refer to these choice problems as the FOSD tasks, which are denoted as π^r and π^c depending on whether the row or column lottery is dominant. The following corollary trivially follows from Proposition 1.

Corollary 1. *If* Π *admits a correlation-sensitive representation as given in the expression* (2.1)*, the following holds for all* $h, m, l \in \mathbb{R}$ *such that* $h > m > l$*.*

1. $\phi(h,l) > \phi(h,m) + \phi(m,l) \iff \pi^{sml} \in \hat{\Pi}$;
2. $\phi(h,l) < \phi(h,m) + \phi(m,l) \iff \overline{\pi}^{sml} \in \hat{\Pi}$;
3. $\phi(h,l) = \phi(h,m) + \phi(m,l) \iff \pi^{sml} \in \Pi,\ \overline{\pi}^{sml} \in \Pi \iff \pi^r \in \hat{\Pi},\ \overline{\pi}^c \in \hat{\Pi},\ \forall \epsilon > 0.$

The corollary states that any correlation-sensitive decision maker has a strict preference for either the row or the column lottery of π^{sml}. That is, in an experimental setting, any decision maker satisfying ISPD must express a preference for the row lottery. By contrast, a decision maker satisfying DSPD must express a preference for the column lottery. We will therefore refer to the row lottery as the ISPD lottery, and to the column lottery as the DSPD lottery. The SML task provides a stringent test of correlation-sensitive preferences in the following sense. Should a decision maker fail to express a preference for the ISPD (DSPD) lottery, it can be concluded that the decision maker will not display increasing ISPD (DSPD) for any choice task.

A correlation insensitive decision maker is indifferent between the row and the column lottery of π^{sml}. If experimental subjects are forced to choose either the row or the column lottery, indifference implies that they choose either option with 50% probability. Subjects satisfying ISPD (DSPD), on the other hand will, safe for decision error, choose the row (the column) lottery. Therefore, a single choice of a single subject cannot be used to infer correlation insensitivity. However, averaging over

π^r	$h(1/3)$	$m(1/3)$	$l(1/3)$
$h+\epsilon(1/3)$	0	0	1/3
$m+\epsilon(1/3)$	1/3	0	0
$l+\epsilon(1/3)$	0	1/3	0

(i) The raw lottery being FOSD

π^c	$h+\epsilon(1/3)$	$m+\epsilon(1/3)$	$l+\epsilon(1/3)$
$h(1/3)$	0	0	1/3
$m(1/3)$	1/3	0	0
$l(1/3)$	0	1/3	0

(ii) The column lottery being FOSD

Table 2.3 Tests of correlation sensitivity

choices, either across different subjects or across different choices within the same subjects, allows to cleanly distinguish between the three categories of preferences we introduced previously.

To gauge the strength of correlation sensitivity, we further consider the FOSD tasks. Correlation insensitivity implies a strict preference for the dominant lotteries under the joint distributions π^r as well as π^c. The SML task should reveal correlation sensitivity even if it is only of second-order importance in the sense of lexicographic preferences. For the FOSD tasks, however, experimental subjects have to violate FOSD in order to express their correlation-sensitive preferences.

The SML task also is a test of transitivity. Because correlation sensitivity is equivalent to intransitivity, Proposition 1 in conjunction with Corollary 2.3.1 imply that a strict preference for the row or the column lottery is equivalent to violating transitivity in Lanzani's framework. However, the SML task provides a test of transitivity in a more general sense. As Lemma A.1 used in the proof of Proposition 1 shows, one needs only impose the completeness axiom, and neither strong independence, nor Archimedean continuity, for a strict preference for the row or the column lottery of the SML task to constitute a violation of transitivity. Note however, when only completeness is imposed, correlation insensitivity is no longer equivalent to transitivity, meaning that transitivity can be violated in ways unrelated to correlation sensitivity. The SML task therefore allows to unambiguously distinguish between transitive and intransitive preferences only within Lanzani's framework, but not in the more general case in which only completeness is imposed.

2.3.2 A comparison to past studies

In this section, we compare the SML task to previous experimental tasks used to test for correlation sensitivity. The approach closest the SML task is testing for correlation sensitivity by manipulating the joint distribution for choices between lotteries with different marginal distributions (Starmer and Sugden, 1993; Humphrey, 1995; Ostermair, 2021; Loewenfeld and Zheng, 2024). If no evidence of correlation sensitivity is observed in these studies, it does not necessarily imply that subjects are not correlation-sensitive. To clarify this point, consider the example task of Starmer and Sugden (1993) illustrated in Table 2.4 below. Subjects have to choose between the row and the column lotteries. The marginal distribution of the row lottery is relatively riskier but has a higher expected value. Note that the choice on the right hand side of the table is the same in terms of the marginal distributions of the lotteries. However, the joint distributions differ, with the lotteries on the left-hand panel being more negatively correlated than on the right-hand panel.

Consequently, observing a subject express a preference for the raw lottery under the joint distribution π and a preference for the column lottery under the joint distribution π' implies that her preferences are characterized by ISPD. The reverse choice pattern implies DSPD. However,

these choice patterns are not implied by increasing or decreasing sensitivity to payoff differences. Formally, $\pi \in \Pi$ and $\overline{\pi}' \in \Pi \Rightarrow \phi(h,l) \geq \phi(h,m) + \phi(m,l)$, and $\pi' \in \Pi$ and $\overline{\pi} \in \Pi \Rightarrow \phi(h,l) \leq \phi(h,m) + \phi(m,l)$, but the reverse direction ($\Leftarrow$) does not hold. Thus, failing to observe these patterns cannot be taken to imply that preferences are correlation-insensitive.

Confronting subjects with this kind of tasks, Starmer and Sugden (1993) find no evidence for correlation-sensitive preferences. As the authors themselves conclude, the implications of these results are unclear as it is possible that correlation sensitivity would be more prevalent in another setting (Starmer and Sugden, 1993, p.253). Intuitively, correlation-sensitive preferences do not necessarily need to manifest in the setting of Starmer and Sugden (1993) if preferences over the marginal distributions of the lotteries are strong enough. For instance, since the row lottery in Table 2.4 yields a higher expected payoff, a given subject might choose it under both correlation structures, even if the very same subject would display correlation sensitivity in another setting. The SML task does away with such ambiguities. Intuitively, since choices are between two lotteries with the same marginal distribution, preferences over marginal distributions cannot play any role. This forces decision makers to reveal their correlation sensitivity.

A second set of studies uses choice triples to test for violations of transitivity as implied by ISPD (Loomes et al., 1991; Baillon et al., 2015). Consider the example displayed in Table 2.5. Subjects choose between the three marginal distributions $A = (8, 0.6; 0, 0.4)$, $B = (18, 0.3; 0, 0.7)$, and $C = (4, 1)$. The cycle $A \succ B$, $B \succ C$, and $C \succ A$ is consonant with ISPD, whereas the reverse cycle $B \succ A$, $C \succ B$, $A \succ C$ is consistent with DSPD. In the absence of decision noise, observing an ISPD (DSPD) conform cycle implies ISPD (DSPD) of preferences. However, following a similar argument as above, not observing such cycles does not imply CSPD. Decision noise, which might affect all three choices required by experimental subjects, further complicates inference about the transitivity of preferences in non-trivial ways (Sopher and Gigliotti, 1993; Regenwetter et al., 2011; Loomes, 2005; Baillon et al., 2015). As the SML task consists of a single choice, it avoids such issues. Apart from these theoretical properties, the SML task has a number of additional advantages from an experimental viewpoint. First, all joint payoff realizations have equal probability, which renders the task easy to understand. Second, our task naturally controls for event-splitting effects that are present in a number of studies on correlation sensitivity (e.g., Loomes, 1988b; Frydman and Mormann, 2018; Bruhin et al., 2022). Third, we avoid displaying duplicated states to experimental subjects, which is common in studies testing for correlation sensitivity while controlling for event-splitting effects (e.g., Loomes et al., 1991; Starmer and Sugden, 1993; Humphrey, 1995; Ostermair, 2021; Loewenfeld and Zheng, 2024). We also avoid "null states" or other states that should be edited out. While it might be argued that these features are undesirable since they could have

π	7 (55%)	0 (45%)
11 (45%)	0%	45%
0 (55%)	55%	0%

(i) Negative correlation structure

π'	7 (55%)	0 (45%)
11 (45%)	45%	0%
0 (55%)	10%	45%

(ii) Positive correlation structure

Table 2.4 An example task from Starmer and Sugden (1993)

π^A	8 (60%)	0 (40%)
18 (30%)	30%	0%
0 (70%)	30%	40%

(i) Lottery A vs lottery B

π^B	18 (30%)	0 (70%)
4 (100%)	30%	70%

(ii) Lottery B vs lottery C

π^C	4 (100%)
8 (60%)	60%
0 (40%)	40%

(iii) Lottery A vs lottery C

Table 2.5 An example task from Loomes et al. (1991)

unexpected effects on behavior, they are present in many of the choice tasks that have been used to test for correlation-sensitive preferences (Loomes, 1988b; Starmer and Sugden, 1993; Humphrey, 1995; Dertwinkel-Kalt and Köster, 2019; Ostermair, 2021; Loewenfeld and Zheng, 2024).

2.3.3 Probing the psychological foundations of correlation sensitivity

Although regret and salience theory both induce ISPD, they build on distinct psychological mechanisms. In both regret and salience theory, decision makers compare payoffs within states of nature, although for different reasons. In regret theory, within-state comparisons of payoffs are deliberate as they impact the decision maker's utility. In salience theory, within-state comparisons of payoffs can be seen as incidental. They impact the decision maker's perception of the choice task. However, there is nothing about within-state difference of payoffs that the decision maker intrinsically values. This has led to the suggestion that payoff comparisons need not necessarily be determined by which payoffs realize jointly but by how decision problems are framed when presented to decision makers (Dertwinkel-Kalt and Köster, 2015; Leland et al., 2019). This approach seems broadly consistent with the description of salience as "a property of states of nature that depends on the lottery payoffs that occur in each state, as they are presented to the decision maker" (Bordalo et al., 2012, p.1256), and the recurring allusions to salience-driven framing effects throughout the paper.

We use this distinction to experimentally differentiate between correlation sensitivity that arises from intentional comparisons of jointly realizing payoffs and correlation sensitivity that arises from payoff comparisons due to the visual presentation of choice problems. We implement a set of three between-subjects treatments as illustrated in Table 2.6. In all three treatments, subjects face a binary choice similar to our SML task in Table 2.1. Following the experimental literature, we describe the joint distribution of lotteries by referring to the underlying states of nature. That is, the payoff generated by the lotteries depends on the realization of state of nature, represented by different fields of a wheel of fortune. In panel (2.6i), which represents the baseline treatment, each column describes the realizations of both lotteries in a given state of the world. In this treatment, correlation-sensitivity can arise because subjects deliberately compare payoffs state-by-state, but it could also arise if subjects incidentally compare payoffs column-by-column. Evidently, it is not possible to distinguish between the two. This display follows previous studies on correlation sensitivity (Starmer and Sugden, 1993; Humphrey, 1995; Ostermair, 2021; Loewenfeld and Zheng, 2024).

The remaining two treatments allow to distinguish between correlation sensitivity arising from state-by-state and column-by-column comparisons of payoffs. Consider the display in the column-

Lotteries	Payoffs & States of nature		
A	h if s_1	m if s_2	l if s_3
B	l if s_1	h if s_2	m if s_3

(i) Baseline treatment

Lotteries	Payoffs & States of nature		
A	h if s_1	m if s_2	l if s_3
B	l if s_3	h if s_1	m if s_2

(ii) Column-effects treatment

Lotteries	Payoffs & States of nature		
A	h if s_1	m if s_2	l if s_3
B	h if s_2	m if s_3	l if s_1

(iii) State-effects treatment

Table 2.6 Between-subjects treatments (with equiproble states and $h > m > l$) for differentiating the distinct mechanisms underlying correlation sensitivity.

effects treatment, illustrated in panel (2.6ii). Note that the two lotteries are perfectly correlated. To a decision maker whose correlation sensitivity is caused by deliberate state-by-state comparisons of payoffs, they are equivalent. However, note that the column-by-column comparison of payoffs is equivalent to that in the baseline treatment. Thus, any correlation sensitivity arising in this treatment can be only attributed to incidental payoff comparisons that arise from presenting payoffs column-by-column. Finally, consider the state-effects treatment illustrated in panel (2.6iii). The payoffs displayed in each column are the same for both lotteries, whereas the joint distribution of payoffs is as in the baseline treatment. Arguably, this choice display allows subjects to readily recognize that both lotteries share the same marginal distribution. Therefore, correlation sensitivity arising in the state-effects treatment cannot be caused by incidental column-by-column comparisons of payoffs, but it can only arise from state-by-state comparisons of payoffs. Arguably, such state-by-state comparisons now require a deliberate effort on behalf of the experimental subjects.

Within the baseline treatment, we further implement a within-subject treatment that varies the timing of outcome feedback. An argument often put forward in the psychology literature is that in order for people to minimize ex-post regret, they have to anticipate immediate outcome feedback (Bell, 1983; Zeelenberg et al., 1996; Zeelenberg, 1999; Somasundaram and Diecidue, 2017). While the timing of outcome feedback is not part of regret theory (Loomes and Sugden, 1982, 1987), we explore the possibility that immediate feedback is important for correlation sensitivity to emerge by providing immediate feedback on some, but not all tasks.[10] In the experiment, participants make a number of choices, one of which is randomly selected to be payoff relevant. When participants receive immediate outcome feedback, they are informed only about the outcome of their choice, but not whether the task was selected for payoff. Note that, when participants do not obtain immediate outcome feedback, they still receive feedback for the payoff relevant task at the end of the experiment. Therefore, while any feedback effects we observe will be ascribed to anticipated regret, it should not be inferred that correlation sensitivity in the absence of immediate feedback cannot be driven by anticipated regret.

[10] Our feedback manipulation is similar to that of Somasundaram and Diecidue (2017), but we are, to the best of our knowledge, the first to test for feedback effects on correlation sensitivity.

2.3.4 Main experimental hypotheses

Testing properties of deterministic models on invariably noisy choice data necessitates imposing assumptions about the nature of noise and probabilistic choice (Loomes and Sugden, 1995; Luce, 1995; Baillon et al., 2015). We impose the minimal assumption that, if individuals are indifferent, they choose at random, and that their probability of choosing a given lottery is (weakly) increasing in their utility of doing so. This implies that correlation-insensitive individuals choose either lottery of the SML task with 50% probability. This random choice benchmark constitutes our null. As regret and salience theory are the main theories implying correlation sensitivity, we derive our alternative hypothesis assuming that preferences are characterized by ISPD, which implies that the ISPD lottery is chosen at a frequency higher than 50%. This motivates Hypothesis 1(a).

Second, if immediate feedback is necessary for correlation sensitivity driven by regret aversion to arise, we might expect the preference for the ISPD lottery to be more pronounced when subjects receive immediate feedback, as opposed to when feedback is only provided at the end of the experiment. This motivates Hypothesis 1(b).

Further, observing choice frequencies different from 50% in the column-effects treatment will provide evidence of correlation sensitivity driven by column-by-column comparisons, whereas observing choice frequencies different from 50% in the states-effects treatment will provide evidence of correlation sensitivity caused by state-by-state comparisons. In the first two of three experiments only the baseline treatment was employed. After having observed choice frequencies lower than 50% in these experiments, we hypothesized, based on introspection, that the observed correlation sensitivity is driven by column-by-column comparisons. This motivates our Hypothesis 1(c) and (d).

Finally, we gauge the strength of correlation sensitivity. If preferences are correlation-insensitive, subjects will, safe for decision noise, choose the dominant lottery in both panels (2.3i) and (2.3ii) of Table 2.3. This implies an overall choice frequency of the row lottery of 50%. Pooling choices for both cases, we can again test for correlation-sensitive preferences by testing whether the ISPD lottery is chosen at an overall frequency higher than 50%.[11] We summarize our experimental hypotheses below.

Hypothesis 1. *Correlation-sensitivity at the aggregate level.*

(a) The ISPD lottery is chosen at a frequency higher than 50%.

(b) The above effect is larger when subjects receive immediate outcome feedback.

(c) The ISPD lottery is chosen at a frequency significantly lower than 50% in the column-effects treatment.

(d) The ISPD lottery is chosen at a frequency insignificantly different from 50% in the state-effects treatment.

[11] An alternative way of testing for correlation-sensitive preferences, that is more in the spirit of Starmer and Sugden (1993), would be to test whether violations of first-order stochastic dominance occur more often when ISPD favors the dominated lottery than when it favors the dominant one.

	Lab	Online 1	Online 2	Online 2	Online 2
Treatment	Baseline	Baseline	Baseline	State	Column
Sample	Students	General	General	General	General
Date	2021.03	2022.06	2022.12	2022.12	2022.12
Tasks	SML	SML	SML	SML	SML
	Event-splitting[a]	FOSD	FOSD	FOSD	FOSD
	Feedback	Feedback	-	-	-
	Attention	Attention	Attention	Attention	Attention
Valid subjects	289	145	158	159	150
Excluded subjects	7	11	59	53	61

Table 2.7 Summary of the experiments

[a] Results from this part are reported in a companion paper (Loewenfeld and Zheng, 2024).

(e) Correlation sensitivity persists even if one lottery is first-order stochastic dominant.

2.4 Experimental procedures

We pre-registered our three experiments with the AEA social science registry under the IDs AEARCTR-0007239, AEARCTR-0009573, and AEARCTR-0010279. Table 2.7 provides a summary of the experiments. The first experiment was conducted in March 2021 at Renmin University of China in Beijing. In June 2022 and December 2022, we conducted two additional online experiments. After excluding subjects who violated a pre-defined attention check, we remain with 289 valid responses from the lab experiment, 145 from the first and 467 from the second online experiment, 158 of which are from the baseline, 159 from the column-effects, and 150 from the state-effects treatment.[12] All participants in the lab experiment were students, whereas only between 17% and 27% of the participants in the online experiments stated to be students, with the majority of the remaining participants being part- or full-time employed. Participants in the lab experiment were around 20 years old on average, while the average age in the online experiments was around 30. 59% of the participants in the lab and the first online experiment were female, whereas around 50% participants of the second online experiment were female in each of the treatments. Additional summary statistics of our samples can be found in the online appendix.

The lab and the first online experiment included SML tasks in the baseline display as well as the immediate feedback treatment. The second online experiment included SML tasks as well FOSD tasks, and consisted of the baseline, column-effects, and state-effects treatment that were described in Table 2.6, but we dropped the immediate feedback treatment.

In the lab experiment, participants completed a total of 35 choice tasks. Among these, 10 choices were between two lotteries with the same marginal distribution.[13] We employed 2 sets of SML tasks (see Table 2.8). Each set consists of three choice tasks with three states and two choice tasks

[12] In all experiments, we included two choice tasks for which one option dominates the other state- and column-wise. It is commonly observed that subjects violate state-wise dominance in both laboratory and online experiments. For example, Samek and Sydnor (2020) documented violations of state-wise dominance ranging from 10% to 60% in the laboratory experiment and from 24% to 50% in the survey experiment. The violation rates observed in our experiments were around 19%. People who were excluded are younger on average, less educated, more likely to be a student, and less likely to be married. However, none of these variables is found to be correlated with correlation sensitivity. Including excluded participants in the analysis does not change our results qualitatively, but reduces effects.

[13] The remaining choice tasks were part of a related study and are described in Loewenfeld and Zheng (2024).

Set 1					Set 2				
Task	a	b	c	d	Task	a	b	c	d
1	73	64	20	-	1	110	33	9	-
2	120	33	0	-	2	101	41	15	-
3	101	53	0	-	3	86	50	3	-
4	149	50	16	0	4	143	32	26	7
5	120	60	20	0	5	94	81	37	13

Table notes: the three-states lotteries always have the following three possible states: $(x^A, x^B) \in \{(c,a),(b,c),(a,b)\}$. The four-state lotteries always have the following four possible states: $(x^A, x^B) \in \{(d,a),(c,d),(b,c),(a,b)\}$. All states are equally likely.

Table 2.8 Parameters for the SML tasks in the lab and first online experiment

with four states, all equiprobable. For one set of SML tasks, subjects did not receive immediate feedback. These choices were presented to subjects in random order among the other 25 choice tasks. The five choices for which subjects received immediate feedback were always encountered at the end of the experiment. After having decided on the choices for which no feedback was provided, subjects were informed that they would make five more decisions for which they would now receive immediate feedback on their choice. Subjects were then exposed to these choice tasks in random order. We chose this particular order so as to avoid potential effects of past feedback on choice tasks for which subjects did not receive feedback. We counterbalanced whether subjects received feedback for set 1 or 2.

The main motivation for the first online experiment was to test for correlation-sensitive preferences using a more general population.[14] The experiment was similar in design to the lab experiment, but it only included SML tasks. In addition, we included four FOSD tasks and one state-wise dominant lottery as an attention check, which, for ease of exposition, we describe this in detail in the online appendix.[15] Subjects in the first online experiment made a total of 19 choices.

The goals of the second online experiment were to disentangle between correlation sensitivity caused by state-by-state and column-by-column comparisons, and to gauge the strength of correlation sensitivity by including FOSD tasks. We implemented all three between-subject treatments discussed above. Each subject encountered all 9 SML tasks with parameters displayed in Table 2.9. We slightly changed the set of parameter values in order to include 3 choice tasks with 6 states.[16] The parameters of the three- and four-state choice tasks used in the second online experiment are shared among all experiments, which ensures comparability. We obtain pairs of first-order stochastic dominant and dominated lotteries by adding a premium of 1, 3, or 9 (approximately 2%, 6%, and 18% of the lotteries' expected value) to either the DSPD or the ISPD lottery, as illustrated in Table 2.3. Each subject further encountered each of the 9 lotteries with one of the

[14] Renmin University of China is generally considered as one of the Chinese top universities and its students are highly trained in mathematics, which might reduce the scope to document correlation-sensitive preferences.

[15] Overall choice patterns are similar to those observed in the second online experiment. For ease of exposition and because we systematically vary premiums in the second online experiment, we focus the discussion on FOSD tasks from this experiment. We also varied the choice display of the FOSD tasks systematically in a way that increases the complexity of the choice tasks to account for the possibility that effects as prescribed by salience theory might only become apparent when choices are sufficiently complex. Results are qualitatively similar.

[16] The inclusion of six-state choice tasks was motivated by an argument that correlation effects as prescribed in salience theory might not be apparent in three-state choice tasks because it is too obvious that both lotteries have the same marginal distribution.

Task	a	b	c	d	e	f
1	73	64	20	-	-	-
2	101	53	0	-	-	-
3	110	22	9	-	-	-
4	149	50	16	0	-	-
5	120	60	20	0	-	-
6	94	81	37	13	-	-
7	93	75	57	39	21	3
8	135	72	50	37	24	8
9	115	75	61	39	27	14

Table notes: the joint distribution of the choice tasks is always given as follows. For the three-state lotteries: $\{(a,c),1/3;(b,a),1/3;(c,b),1/3\}$. The four-state lotteries always have the following four possible states: $\{(a,d),1/4;(b,a),1/4;(c,b),1/4;(d,c),1/4\}$. The six-state lotteries always have the following four possible states: $\{(a,f),(b,a),(c,b),(d,c),(e,d),(f,e)\}$. All states are equally likely.

Table 2.9 Parameters for the SML tasks in the second online experiment

three premiums. That is, they encountered each lottery with a premium of 1, 3, or 9.[17] Premiums are varied between subjects such that the same number of subjects encounter a given parameter set for a given premium. As an attention check, we also included two choices for which one lottery dominates the other state- and column-wise. Subjects made a total of 29 $(3\times 9+2)$ lottery choices.

In all experiments, participants read that for each choice task, there were two options with payoffs that depend on the turn of a personal wheel of fortune.[18] In all treatments, choice problems were displayed to subjects as shown in Figure 2.1a. In the baseline treatment of the lab and first online experiment, the wheel of fortune was described as having 99 equally likely fields. The implementation of the column-effects and the state-effects treatment necessitated a slightly different display. To avoid overloading choice presentation, we decided to implement the state space through a wheel of fortune with up to six equiprobable fields of different colors. We color-coded the fields to improve state-by-state comparability. See Figure 2.1b-2.1d for examples. The implementation of the baseline treatment ensures comparability between the different experiments. Before they were allowed to start real choice tasks, subjects had to answer a set of comprehension questions correctly. In case a subject gave a wrong answer, they received feedback intended to help them understand the task at hand.

During the experiment, payoffs were displayed in an experimental currency that was translated into Yuan at a rate of 0.5 in the lab and at a rate of 0.4 in the online experiments. To avoid any unwanted effects of the way in which choices are presented, we randomized the order in which states appear. Lotteries were referred to in neutral language, as "Option A" and "Option B". We also randomized which lottery was labelled option A and B. All of this randomization was done at the subject level. After the choice tasks, subjects were prompted to answer a short questionnaire. Upon finishing the experiment, subjects received their payment. Participants in the lab study received a show-up fee of 10 Yuan and had one randomly selected choice paid out. Participants in the online experiments received a participation fee of 9 Yuan and had a 1/3 chance of having one of their choices paid out. Subjects received an average payoff of around 41 Yuan in the lab

[17] The premium of 1 is chosen because it is the smallest possible premium while sticking to integer values. We then increase the premiums by a factor of three.

[18] By referring to a *personal* wheel of fortune, we address the concern that correlation structure might impact subjects' choices because of other-regarding preferences.

	Fields 1-33 (33.3%)	Fields 34-66 (33.3%)	Fields 67-99 (33.3%)
○ Option A	101	53	0
○ Option B	0	101	53

(a) Lab and 1st online experiment: Baseline

○ Option A	**101** if **Red**	**53** if **Blue**	**0** if **Green**
○ Option B	**0** if **Red**	**101** if **Blue**	**53** if **Green**

(b) 2nd online experiment: Baseline

○ Option A	**101** if **Red**	**53** if **Blue**	**0** if **Green**
○ Option B	**0** if **Green**	**101** if **Red**	**53** if **Blue**

(c) 2nd online experiment: Column-effects

○ Option A	**101** if **Red**	**53** if **Blue**	**0** if **Green**
○ Option B	**101** if **Blue**	**53** if **Green**	**0** if **Red**

(d) 2nd online experiment: State-effects

Figure 2.1 The display of choice problems in different treatments and experiments

experiment, 17 Yuan in the first and 16 Yuan in the second online experiment. The lab experiment lasted around 30 minutes, and the two online experiments took between 10 and 15 minutes. All experiments were programmed with oTree (Chen et al., 2016). The complete instructions can be found here.

2.5 Results

2.5.1 Correlation sensitivity at the aggregate level

We begin by testing for correlation sensitivity. At this stage, we combine data from choices with and without immediate feedback. We analyze data from SML tasks from all three experiments for the baseline treatment, resulting in a sample of 5762 choices made by 592 participants. Participants in the lab experiment chose the ISPD lottery with a frequency of 48.1%, while participants in the first online experiment chose it with a frequency of 45.1%, and subjects in the second online experiment chose it with a frequency of 39.0% (see Figure 2.2). As preregistered, we test hypothesis 1(a) by estimating logistic models that regress a variable indicating whether a subject chose the ISPD lottery on a constant. This approach allows to conveniently account for repeated observations from the same subject by clustering standard errors at the individual level. We find that the choice frequency differed only marginally from the 50% random-choice benchmark for the lab experiment ($p = 0.08$), but significantly in the first and second online experiments ($p = 0.003$ and $p < 0.001$, respectively). Correlation sensitivity appears to be insignificant for the lab experiment participants, slightly stronger for the first online experiment, and even stronger for the second online experiment. As we discuss in more detail in the online appendix, these disparities may be due to a combination of differences in the subject pool, changes in the choice display, as well as the number of states of the choice tasks. Notably, correlation sensitivity seems to be particularly evident for choice tasks with six states. Overall, the results provide strong evidence against Hypothesis 1(a). In contrast to the predictions of regret and salience theory, the aggregate choices suggest a modest preference for the DSPD lottery.

Given that the choices of subjects in our lab studies are close to random, one might question the robustness of our results. In section 2.6, we discuss additional results from two recent pilot

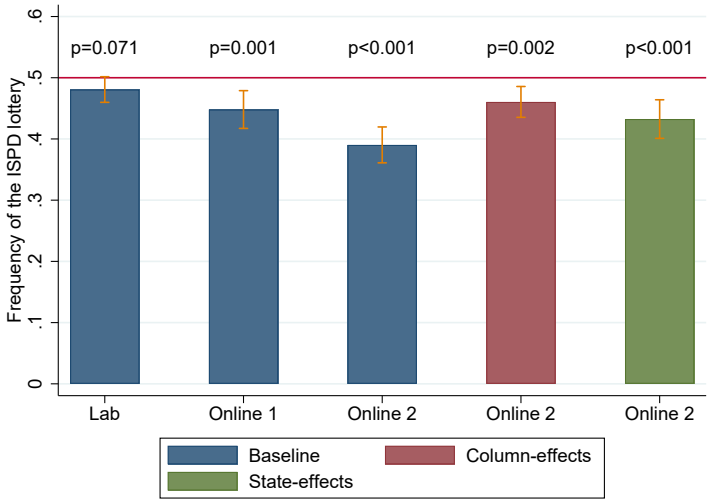


Figure 2.2 Choice frequencies of the ISPD lottery

studies with Chinese University students and an unrelated experiment with French students that alleviate such concerns. We further discuss exploratory analysis into whether other characteristics of lotteries such as their skewness or relative skewness make them more conducive to correlation-sensitivity, but fail to find robust patterns. We provide choice frequencies for each lottery task in the online appendix. In the next step, we test for the impact of immediate feedback. We find that choices were not significantly influenced by immediate feedback in both the lab and the first online experiment. In the lab experiment, 47.8% of choices were for the ISPD lottery when subjects received immediate feedback, while 48.4% were without feedback. Similarly, in the online experiment, subjects chose the ISPD lottery at a frequency of 44.8% with immediate feedback and at a frequency of 45.4% without it. As preregistered, we run logistic regressions with a dummy variable for ISPD lottery choice as the dependent variable and a dummy variable indicating whether immediate feedback was provided as the explanatory variable. We find that feedback did not significantly impact choices in either the lab ($p = 0.73$) or the online experiment ($p = 0.83$), or when we pool the two ($p = 0.68$).[19] To assess the precision of these null results, we calculate 95% confidence intervals for the effect of feedback on choice frequencies. The confidence intervals for the immediate feedback effect are [-0.046, 0.032] for the lab experiment, [-0.046, 0.032] for the online experiment, and [-0.037, 0.024] for the pooled sample. This suggests that the null result on feedback effects is precisely estimated and not the result of noisy data. Therefore, we reject Hypothesis 1(b). In the online appendix, we explore heterogeneity in feedback effects and find that participants reporting a higher tendency to feel regret in a hypothetical investment scenario (Guiso, 2015) chose the ISPD lottery less often under immediate feedback than without feedback. We interpret this finding as suggestive evidence that anticipated feedback might play an important role only for individuals with a high propensity to experience regret.

We next turn to the column-effects and state-effects treatments to shed some light on the drivers

[19] Regressions results can be found in the online appendix. Unless otherwise noted, p-values are obtained from Wald Chi-Square test with standard errors clustered at the subject level. We also obtained similar results using pre-registered non-parametric Wilcoxon signed-rank tests (p-values of 0.58, 0.22, and 0.20 for the lab, online, and pooled samples, respectively).

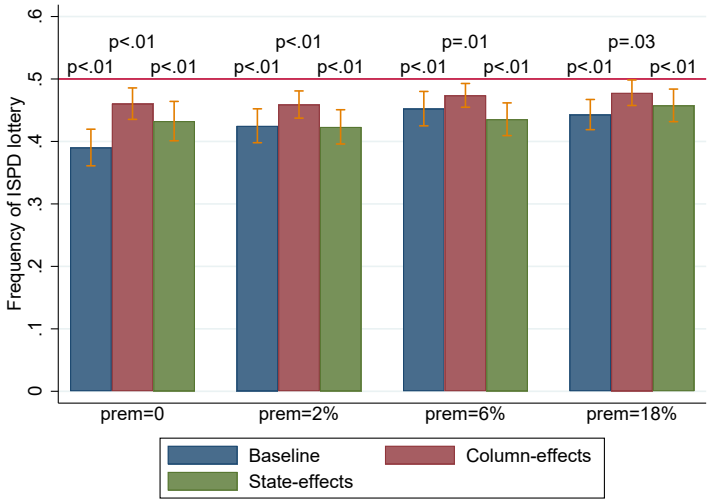


Figure 2.3 Choice frequencies of the ISPD lottery by levels of premium

of correlation-sensitivity. In the column-effects treatment, participants chose the ISPD lottery with a frequency of 46.1%, while in the state-effects treatment, they chose it with a frequency of 43.3% (see Figure 2.2). As preregistered, we again use logistic regressions with standard errors clustered at the subject level (see the online appendix for more details), and find that the choice frequency of the ISPD lottery is significantly below 50% in both treatments ($p = 0.002$ in the column-effects treatment and $p < 0.001$ in the state-effects treatment). The logistic regressions also suggest that the difference in choice frequencies of 7.1 percentage points (ppt) between the baseline treatment and the column-effects treatment is statistically significant ($p < 0.001$). The difference of 4.1 ppt between the baseline and the state-effects treatment is marginally significant at $p = 0.054$, and the 2.8 ppt difference between the column-effects and the state-effects treatment is not statistically significant ($p = 0.17$, based on a Chi-squared test).

We do not reject Hypothesis 1(c) but reject Hypothesis 1(d). If anything, state-by-state comparisons seem to be somewhat more important, although the difference between the state-effects and the column-effects treatment is not statistically significant. The findings suggest that correlation sensitivity in the baseline treatment arises from both deliberate state-by-state comparisons of payoffs, as well as incidental column-by-column comparisons of payoffs. Importantly, results from both treatments again imply modest aggregate correlation sensitivity in line with DSPD.

The results so far provide very consistent evidence for DSPD. However, since the effects were observed with same marginal lotteries, these could, in principle, be of second-order importance only. To test the robustness of the observed correlation sensitivity, we now turn to the FOSD tasks. Figure (2.3) displays the choice frequencies of the ISPD lottery as a function of the payoff premium. As can be seen, the choice frequencies are always significantly smaller than 50% at the 5% level in all three treatments, even for the highest level of premium, but seem to be moving closer to 50% as the premium is increased. To test for the impact of the size of the premium on choice behavior more formally, we run the following, preregistered, logistic regressions separately

	(1) Baseline	(2) Column-effects	(3) State-effects
Variables	ISDP	ISDP	ISDP
p2	0.144**	-0.006	-0.038
	(0.070)	(0.062)	(0.063)
p6	0.256***	0.053	0.012
	(0.071)	(0.060)	(0.064)
p18	0.217***	0.070	0.102
	(0.073)	(0.064)	(0.065)
Constant	-0.446***	-0.158***	-0.271***
	(0.063)	(0.052)	(0.066)
Observations	4,266	4,293	4,050
Individuals	158	159	150

Table notes: Significance levels are: * for p≤0.1; ** for p≤0.05; *** for p≤0.01.

Table 2.10 Logistic regressions on SML tasks

for each treatment.

$$ISDP_{i,t} = c + \beta_1 p2_{i,t} + \beta_2 p6_{i,t} + \beta_3 p18_{i,t} + \epsilon_{i,t}, \tag{2.2}$$

where $p2_{i,t}$, $p6_{i,t}$, and $p18_{i,t}$, with i being the index of subjects and t being the index of treatments, are dummy variables that indicate the levels of premium 2%, 6%, and 18% respectively. Zero premium is the omitted category. Table 2.10 reports the regression results. Higher premiums significantly reduce DSPD only in the baseline treatment, but not in the other two treatments.

Another way to look at the effects of correlation sensitivity is to consider its impact on rates of violations of first-order stochastic dominance. Pooling all levels of the payoff premium, subjects in the baseline treatment violate first-order stochastic dominance at a rate of 15.2% when the DSPD lottery is dominant. The rate of FOSD violations is increased by 78% to 27.1% when it is the ISPD lottery that is dominant. In the column-effects treatment, the rate of FOSD violations is 9.5% when the DSPD lottery is dominant. This rate is increased by 62% to 15.4% when it is the ISPD lottery that is first-order stochastic dominant. In the state-effects treatment, subjects violate FOSD at a rate of 12.9% when the DSPD lottery is dominant, but at a 95% higher rate of 25.1% when the ISPD lottery is dominant. For all comparisons, the increase in the rate of FOSD violations is statistically significant ($p < 0.001$, logistic regression with standard errors clustered at the subject level).

Overall, the results from FOSD tasks provide further evidence of a DSPD effect of modest size. FOSD is much more predictive of aggregate choice patterns than the joint distribution of lotteries. However, our results also suggest that correlation sensitivity is responsible for a sizable fraction of the FOSD violations we observe. This suggests that correlation sensitivity is not only of second-order importance but can exert a significant influence over participants' choices. We do not reject Hypothesis 1(e).

We summarize our findings on correlation sensitivity at the aggregate level below.

Result 1. *On SML tasks and treatment effects:*

(a) In the baseline treatment, we find that subjects chose the ISPD lottery at a frequency of 48.1% in the lab experiment, 45.1% in the first online experiment, and 39.0% in the second online

experiment. The choice frequencies differ from the 50% random choice benchmark marginally in the lab experiment ($p = 0.08$) and significantly in both online experiments ($p < 0.002$). We reject Hypothesis 1(a).

(b) Estimating the effect of immediate feedback on choices, we find a precisely null effect. We reject Hypothesis 1(b).

(c) The ISPD lottery is chosen at a frequency of 46.1% in the column-effects treatment, which is significantly different from 50% at $p = 0.002$. We do not reject Hypothesis 1(c).

(d) The ISPD lottery is chosen at a frequency of 43.3% in the state-effects treatment, which is significantly different from 50% at $p < 0.001$. We reject Hypothesis 1(d).

(e) In all three treatments, we find evidence for DSPD even when one lottery in a pair is first-order stochastic dominant. Pooling all levels of the payoff premium, we find that the ISDP lottery is chosen at a frequency of 44.0% in the baseline treatment, 47.0% in the column-effects treatment, and 43.9% in the state-effects treatment (all with $p < 0.001$). We do not reject Hypothesis 1(e).

2.5.2 Correlation sensitivity at the individual level

After discussing correlation sensitivity at the aggregate level, we turn to explore individual heterogeneity using latent class analysis based on structural equation models. For this purpose, we focus on our second online experiment, since the inclusion of FOSD tasks allows for a richer analysis. We divide choice tasks into the three different categories displayed in Table 2.3: SML tasks, FOSD tasks where the ISPD lottery is dominant, and FOSD tasks where the DSPD lottery is dominant. For each category, we sum all choices a subject made for the ISPD lottery and specify that the resulting variables are distributed according to a binomial distribution with 9 trials. We estimate latent class models pooling all observations from the different treatments.[20] We estimate latent class models with up to seven classes. For models with more classes, convergence fails. Among the estimated models, we select the one with the lowest Bayesian information criterion (BIC) value, which is a model with four classes. Considering posterior probabilities, i.e., the probability of class membership for each individual, we find that the medium participant is assigned to one class with 95% probability and only about 20% of the participants are assigned to one class with less than 75% probability. This suggests that most subjects can be assigned to one of the classes with high probability. Table 2.11 reports the results from the latent class analysis. The behavior of individuals in Class 1 is nearly perfectly characterized by correlation insensitivity. Individuals in this class choose the ISDP lottery with a frequency that does not differ significantly from the 50% random-choice benchmark for the SML tasks and respect FOSD almost perfectly. That is, the ISPD lottery is chosen practically always when it is dominant and practically never when it is dominated.

[20] As we are able to recover the frequency of each class for each treatment, this approach allows for the possibility that different latent classes emerge in the different treatments while avoiding potential issues of unstable classes that could arise from small samples. This approach also enhances interpretability of our results.

	By tasks	% ISPD (SD)	95% CI	By treatments	% Subjects	Averaged
Class 1	Same marginal	0.484 (0.012)	[0.461 - 0.507]	Baseline	0.335	0.494
	ISPD-FOSD	0.987 (0.005)	[0.977 - 0.998]	Column	0.640	
	DSPD-FOSD	0.023 (0.005)	[0.014 - 0.032]	State	0.505	
Class 2	Same marginal	0.227 (0.024)	[0.180 - 0.273]	Baseline	0.191	0.167
	ISPD-FOSD	0.334 (0.029)	[0.276 - 0.391]	Column	0.094	
	DSPD-FOSD	0.229 (0.024)	[0.181 - 0.277]	State	0.216	
Class 3	Same marginal	0.405 (0.029)	[0.348 - 0.462]	Baseline	0.318	0.207
	ISPD-FOSD	0.774 (0.042)	[0.691 - 0.857]	Column	0.156	
	DSPD-FOSD	0.082 (0.017)	[0.049 - 0.114]	State	0.147	
Class 4	Same marginal	0.507 (0.033)	[0.442 - 0.572]	Baseline	0.156	0.132
	ISPD-FOSD	0.537 (0.033)	[0.472 - 0.601]	Column	0.110	
	DSPD-FOSD	0.445 (0.032)	[0.383 - 0.507]	State	0.131	

Table 2.11 Latent-class analysis

Classes 2 and 3 display correlation sensitivity, both consistent with DSPD. Individuals in class 2 display strong DSPD. They choose the DSPD lottery at a frequency of 77% for the SML tasks, at a frequency of 67% when it is first-order stochastic dominated, and at a frequency of 77% when it is first-order stochastic dominant. Individuals in this category seem to be relatively unaffected by FOSD and seem to implement their correlation sensitivity somewhat imperfectly, especially compared to individuals in class 1 who satisfy CSPD near perfectly. The behavior of individuals in class 3 might be best characterized as weak DSPD. Overall, choice behavior in this class is similar to that in class 1 but is somewhat skewed towards the DSPD lottery. The DSPD lottery is chosen at a frequency of 59% for the SML tasks, 23% when it is first-order stochastic dominated, and 92% when it is first-order stochastic dominant. We posit that choices in class 2 can be interpreted as expressing deliberate correlation sensitivity, whereas choice behavior in class 3 could be interpreted as stemming from individuals who are generally correlation insensitive, but whose choices are somewhat biased towards DSPD.

Finally, the fourth class seems to capture random behavior. For all three types of choice tasks, choice frequencies do not differ significantly from 50%. What is striking is the absence of any class of individuals who display behavior consistent with ISPD. This suggests that ISPD is not only rejected as the property governing aggregate behavior, but that virtually none of our participants display behavior that is characterized by ISPD.[21]

Averaging over all three treatments, 49% of the participants are assigned to the correlation-insensitive class 1, while 17% and 21% are assigned to classes 2 and 3 that capture strong and weak DSPD respectively, and 13% are assigned to the random-choice class 4. This suggests that although correlation insensitivity is the predominant category, a sizable fraction of the participants are characterized by DSPD.

Comparing the fractions of participants assigned to the four classes across treatments reveals interesting differences. In all treatments, a similar fraction of 11%-16% of participants are assigned to the random-choice class 4. It is reassuring that these fractions do not differ greatly between treatments. The fraction of subjects assigned to the correlation-insensitive class 1 is only 34% in the baseline treatment, 51% in the state-effects treatment, and reaches 64% in the column-effects

[21] When estimating a model with 7 latent classes, a class characterized by moderate ISPD emerges. Subjects in this class choose the ISPD lottery at frequency of 67% for the SML tasks, 83% when the ISDP lottery is first-order stochastic dominant, and 58% when it is first-order stochastic dominated. A negligible fraction of 1.4% of the participants are assigned to this class.

treatment. This pattern mirrors our aggregate results, which show that subjects display the highest correlation sensitivity in the baseline treatment, followed by the state-effects and column-effects treatments.

Similar fractions of participants in the baseline and state-effects treatments are assigned to the strongly correlation-sensitive class 2, namely 19% and 22% respectively, while only about 9% of participants in the column-effects treatment are assigned to this class. This suggests that strong and consistent DSPD may be primarily driven by deliberate state-by-state comparisons of payoffs. Finally, about 32% of participants in the baseline treatment are assigned to the moderate-DSPD class 3, whereas the corresponding fractions in the column-effects and state-effects treatments are 16% and 15%. The high prevalence of this class in the baseline treatment may be explained by the fact that both column-by-column and state-by-state comparisons of payoffs are aligned in this treatment. This might make it more challenging to discern that both lotteries share the same marginal distribution or that one lottery is dominant.

2.6 Discussion

Our results provide consistent evidence of modest DSPD on the aggregate. The latent-class analysis suggests that this aggregate effect might be driven by a minority of subjects who display strong DSPD, even if it means violating FOSD. Importantly, the analysis does not produce an equivalent class of participants displaying strong ISPD. These findings are in contrast with previous studies that are based on manipulations of the joint distribution and tend to report null results (Starmer and Sugden, 1993; Humphrey, 1995; Ostermair, 2021; Dertwinkel-Kalt and Köster, 2021; Loewenfeld and Zheng, 2024), or studies using the trade-off method that report evidence for ISPD (Bleichrodt et al., 2010; Baillon et al., 2015). Our ability to uncover evidence for DSPD might be attributed to several factors. The SML task improves on previous tests by being more suitable to correlation-sensitive preferences in a theoretical sense and avoids some features that seem undesirable to an experimenter, such as lack of incentive compatibility or duplicated states.

As our findings clearly reject the hypothesis that behavior is characterized by ISPD, both at the aggregate and individual levels, our results strongly reject both regret and salience theory. We would like to reiterate that it is ISPD that allows the theories to rationalize classical behavioral anomalies, such as the Allais paradox or simultaneous gambling and insurance, as well as the commonly observed preference for right-skewed risks. Broadly speaking, DSPD produces the opposite of the commonly observed patterns, such as preferences for negative skewness and an aversion to long-shot lotteries.

As the observed correlation sensitivity contradicts regret and salience theory, it is important to investigate its drivers. Our three between-subject treatments suggest that both incidental payoff comparisons resulting from the framing of choices and deliberate state-by-state comparisons contribute to the observed correlation sensitivity. The latter suggests that some subjects may have genuine preferences that are impacted by the correlation of payoffs across states. Otherwise, it is difficult to understand why subjects would compare payoffs state-by-state across different columns.

While regret and salience theory postulate an increase in sensitivity to within-state differences, one could also argue that decreasing sensitivity to payoff differences has strong intuitive appeal. In EUT, for instance, decreasing sensitivity to increments of wealth is commonly assumed to explain risk aversion. In prospect theory, decreasing sensitivity to incremental losses is used to explain risk aversion in the gain domain and risk-seeking in the loss domain, as well as probability weighting (Tversky and Kahneman, 1992). Our results suggest that decreasing sensitivity is also the predominant pattern governing within-state comparisons.

We believe that our null effect of immediate outcome feedback does not provide good evidence against the possibility that correlation sensitivity is, at least to some degree, driven by an aversion to regret. Since subjects receive outcome feedback on the payoff relevant lottery in any case at the end of the experiment, it is possible that simply changing the timing of the feedback was not sufficient to alter subjects anticipation of regret. We would also like to highlight that our null finding for immediate feedback effects is compatible with studies that document feedback effects (e.g., Zeelenberg et al., 1996; Zeelenberg, 1999). In these studies choices can usually be used to manipulate the outcome feedback one obtains, which is not possible in our setting. Moreover, ISPD is not necessary to rationalize such feedback effects. All one has to assume is that people feel less regret in the absence of feedback on the forgone outcome, as modelled by Bell (1983), for instance. Applied research on regret has often modeled regret in this vein (see, e.g., Filiz-Ozbay and Ozbay, 2007; Engelbrecht-Wiggans and Katok, 2008; Strack and Viefers, 2021; Zheng, 2021).

We would like to stress that the tendency for individuals to display moderate DSPD seems fairly robust across different populations, presentation formats, and parameters, and is therefore unlikely to be an artifact of our design choices. In addition to the experiments reported in the main part, we recently conducted two additional lab experiments using same-marginal lotteries as part of pilot studies with undergraduate students from Peking University and Southwestern University of Finance and Economics (SWUFE). Only three-state lotteries were used in these studies. The experiment in Peking was fairly similar to our second online experiment while excluding the column-effects treatment. Since it was adapted for eye tracking, the presentation format of the choice tasks changed dramatically.[22] We documented strong DSPD for both the baseline and state-effects treatments. The experiment in SWUFE was similar to our first lab experiment, and strong DSPD was also documented in the loss domain and for mixed gambles.[23] Additional evidence comes from an unrelated experiment on delegated risk taking with undergraduate students in Toulouse, France. As part of the experiment, subjects ($N = 144$) had to choose between two pairs of first-order stochastic dominant and dominated lotteries under different correlation structures. We observed weak and marginally significant DSPD for a lottery pair with three states, and strong

[22] For instance, payoff tables were colored with different shades to represent various states of the world, with payoffs displayed in the center of each cell. Furthermore, these payoff tables were presented in either vertical or horizontal orientation.

[23] For 52 undergraduate students from SWUFE, we observed a choice frequency of 26% for the ISPD lottery when lotteries were in the gain domain. We found no significant difference when the lotteries' payoffs were mixed (29%) or when they were in the loss domain (29%). The 40 undergraduate students from Peking University (China) chose the ISPD lottery at a frequency of 39% and 41% in the baseline and state-effects treatments.

DSPD for a six states-choice task.[24] This pattern is consistent with the finding that correlation sensitivity appears to be more pronounced for higher number of states. Moreover, recall that we document correlation sensitivity even in the state-effect treatment of the second online experiment, in which state-by-state comparisons require subjects to exert cognitive effort. This makes it unlikely that the documented effects are driven by subjects reverting to some rule of thumb because the choice tasks are too cognitively demanding. Finally, as an addition to the first online experiment, we directly test for an impact of increased complexity by artificially splitting all states in two, similar to a manipulation carried out by Leland et al. (2019). As we discuss in more detail in the online appendix, the manipulation increases decision noise but seems to have no effect on correlation-sensitivity. These results further suggest that the documented DSPD is not an artifact of a too-complex decision environment.

Given the increase in complexity that arises from intransitive preferences, should (applied) economists allow for correlation sensitivity in their models? Although we find some evidence for correlation sensitivity, we take our results to suggest that, in most applications, the steep price economists have to pay in added complexity when allowing for correlation sensitivity might not be worth it. We arrive at this conclusion for mainly two reasons. First, correlation-sensitive preferences satisfying DSPD induce, broadly speaking, the opposite of the commonly observed patterns such as skewness seeking. Second, although we do observe evidence for DSPD, the overall effect size on the aggregate is rather small. Considering the baseline treatment of the second online experiment, which is the treatment for which we observe the strongest correlation sensitivity, we find an overall choice frequency of 39% for the ISPD lottery in the SML tasks. Using Cohen's g as a rough measure of effect size, this constitutes a small effect.[25] It can be argued that the ability of correlation-sensitive preferences to rationalize commonly observed behavioral patterns results largely from its capability to endogenize the probability weighting of cumulative prospect theory (Tversky and Kahneman, 1992). The recent literature on behavioral inattention and Bayesian updating (Gabaix, 2014; Enke and Graeber, 2021) may provide a way forward without violating transitivity.

Finally, we would like to stress that our study examines correlation-sensitivity in the important but specific setting in which choices are made in a static setting. Comparisons of joint payoff realizations might induce behavior to be more strongly correlation-sensitive in other settings. For instance, Loewenfeld (2024) studies a setting of delegated risk-taking in which principals can reward agents ex-post. He shows that a tendency of principals to condition bonus payments on an ex-post comparison of the realized outcomes can render bonus payments, and therefore agents' incentives to choose between different actions, strongly correlation-sensitive. In a similar vein, studies in the cognitive psychology literature show that an ex-post comparison of the realized and forgone outcome can lead to subjects experiencing regret, which can influence future choices (Camille et al.,

[24] For the three-states choice tasks, 90% of the principals choose the dominant lottery when it is favored by DSPD, and 84% choose it when it is favored by ISDP ($p = 0.095$, McNemar's test). For the six state choice task, 87% of the principals choose the dominant lottery when it is favored by DSPD, but only 59% choose it when it is favored by ISPD ($p < 0.001$, McNemar's test).

[25] If p is the choice frequency, Cohen's g is calculated as $g = |0.5-p| \cdot g < 0.05$ is categorized as negligible, $g \in [0.05, 0.15)$ as small, $g \in [0.15, 0.25)$ as medium, and $g > 0.25$ as large.

2004; Coricelli et al., 2007). It seems conceivable that this mechanism might induce correlation-sensitive behavior in settings in which decision makers have to learn from their choices (Hertwig et al., 2004; Hart and Mas-Colell, 2000; Hart, 2005). The theoretical literature on correlation-sensitive preferences (Loomes and Sugden, 1982, 1987; Bordalo et al., 2012; Lanzani, 2022) will provide valuable guidance in tackling this important issue.

2.7 Conclusion

In this paper, we proposed a theory-tailored experimental task, namely the SML task, to test for correlation-sensitive preferences in risk-taking (Lanzani, 2022). To assess the strength of these preferences and understand their drivers, we further introduced different treatments built on the SML task. In a series of experiments with over 900 participants, we found that aggregate choices displayed modest evidence for decreasing sensitivity to differences in jointly realized payoffs (i.e., DSPD), which contrasts with what regret and salience theory have advocated. Not only could the documented correlation-sensitive preferences survive in face of first-order stochastic dominance, but they were also robust to the absence of immediate outcome feedback. Additionally, we found that both column-by-column and state-by-state payoff comparisons seemed to impact decision-making, but the latter played a more critical role, suggesting that correlation sensitivity arises mainly due to deliberate considerations such as regret avoidance rather than incidental biases such as salience. Finally, our latent-class analysis discovered that only a relatively small sample of subjects contributed to the documented evidence for correlation sensitivity.

Chapter 3

Salience or event-splitting? An experimental investigation of correlation sensitivity in risk-taking

Moritz Loewenfeld[1] Jiakun Zheng[2]

Abstract

Salience theory relies on the assumption that not only the marginal distribution of lotteries, but also the correlation of payoffs across states impacts choices. Recent experimental studies on salience theory seem to provide evidence in favor of such correlation effects. However, these studies fail to control for event-splitting effects (ESE). In this paper, we seek to disentangle the role of correlation and event-splitting in two settings: 1) the common consequence Allais paradox as studied by Bordalo et al. (2012), Frydman and Mormann (2018), and Bruhin et al. (2022); 2) choices between Mao pairs as studied by Dertwinkel-Kalt and Köster (2019). In both settings, we find evidence suggesting that recent findings supporting correlation effects are largely driven by ESE. Once controlling for ESE, we find no consistent evidence for correlation effects. Our results thus shed doubt on the validity of salience theory in describing risky behavior.

Keywords: Salience; Event-splitting; Probability weighting; Concordance; Experiment

JEL Classifications: C91; D81; D91

[1] Toulouse School of Economics, University Toulouse Capitole, 1 Esplanade de l'Université, Toulouse 31080 Cedex 06, France. Email(✉): moritz.loewenfeld@tse-fr.eu.
[2] Aix Marseille Univ, CNRS, AMSE, Centrale Marseille, Marseille, France. Email (✉): jiakun.zheng@outlook.com.

3.1 Introduction

Due to its strong psychological appeal and its ability to rationalize behavior in such diverse areas as finance, industrial organization, advertising, and politics, salience theory (Bordalo et al., 2012) has become increasingly popular in recent years.[3] Salience theory builds on the premise that outcome comparisons within states of nature are an important driver of decision making under risk. In particular, states with a higher outcome contrast attract the decision maker's attention and receive greater decision weights. This assumption implies that decisions are not only driven by the marginal distributions of the lotteries, but also by the correlation of payoffs across states. A few recent experimental studies aim to test the key assumption of salience theory, and report correlation effects as predicted by salience theory (Bordalo et al., 2012; Frydman and Mormann, 2018; Dertwinkel-Kalt and Köster, 2019; Bruhin et al., 2022).

However, a potential issue arises due to the experimental design of these studies. In the experiments, participants had to choose between two lotteries. Participants were presented with these choices under different correlation structures. However, as the correlation of outcomes changed, so did the number of states that was displayed to subjects. Under salience theory such changes to the visual presentation of the choice task should not impact behavior. However, in the literature of testing regret theory (Loomes and Sugden, 1982; Bell, 1982), which is very similar to salience theory (Herweg and Müller, 2021; Lanzani, 2022), Starmer and Sugden (1993) provide evidence that such seemingly benign changes in the presentation of the choice problem can have a tremendous effect on decisions, even when the correlation of payoffs remains constant. Following Starmer and Sugden (1993), we refer to such effects as Event-splitting-Effects (ESE). Any correlation effects claimed in recent studies on salience theory could be attributed to ESE.

In this paper, we aim to disentangle correlation effects from ESE in settings considered in recent studies on salience theory. We take up key aspects of Bordalo et al. (2012), Frydman and Mormann (2018) and Bruhin et al. (2022) who study correlation effects in the context of the common consequence Allais paradox (Allais, 1953). We also consider the experimental set-up adopted by Dertwinkel-Kalt and Köster (2019) who use specific lottery pairs - Mao pairs (Mao, 1970).[4] In order to cleanly disentangle ESE and correlation effects, we implement a between-subject design inspired by Starmer and Sugden (1993). For subjects in the replication treatment, correlation effects and ESE are introduced at the same time. For subjects in the control group, ESE are well controlled for, which allows observing pure correlation effects.

Our findings suggest that the effects ascribed to changes in correlation described in recent studies (Bordalo et al., 2012; Bruhin et al., 2022; Frydman and Mormann, 2018; Dertwinkel-Kalt and Köster, 2019) might be attributable changes in the choice display, rather than changes in the correlation structure. Once we control for ESE, we find no evidence for systematic correlation effects. The ESE we document can be mostly explained by a tendency to overweight payoffs that are displayed in multiple states, even if the probability of the payoffs remains constant.

[3] See Bordalo et al. (2022) for a comprehensive overview on this growing literature.
[4] Mao pairs consist of two lotteries that share the same expected value, and the same variance, but the skewness of one lottery is the negative of the skewness of the other lottery.

Our paper naturally relates to the literature on salience and regret theory. Within the framework of salience theory ((Bordalo et al., 2012; Dertwinkel-Kalt and Köster, 2019), joint realizations characterized by substantial differences in payoffs tend to capture more attention due to their heightened salience, consequently leading to an uneven distribution of decision weight. Regret theory, on the other hand, operates on the premise that a decision maker's utility is shaped by the comparative analysis of payoffs resulting from different choices. When a decision maker comes to realize that an alternative choice could have yielded a superior payoff, a sensation of regret is triggered (Loomes and Sugden, 1982; Bell, 1982). Despite the distinctly different ideas behind the two theories, Herweg and Müller (2021) show that in binary choice situations, regret theory, as proposed by Loomes and Sugden (1982), is a special case of salience theory (Bordalo et al., 2012), which in turn is a special case of generalized regret theory (Loomes and Sugden, 1987). Within the specific framework of our experiments, participants were tasked with making binary choices. This particular experimental setup aligns seamlessly with the decision paradigm outlined by Herweg and Müller (2021) and Lanzani (2022). Consequently, their assertion that both theories yield congruent predictions remains applicable.[5]

Despite the lively interest in regret and salience theory, the current experimental evidence remains inconclusive. Starmer and Sugden (1993) showed that the results of a series of earlier publications that seemed to provide evidence for correlation effects as predicted by regret theory were confounded by ESE, similar to the more recent studies on salience theory (Bordalo et al., 2012; Bruhin et al., 2022; Frydman and Mormann, 2018; Dertwinkel-Kalt and Köster, 2019). Once controlling for ESE, Starmer and Sugden (1993) report that correlation effects as predicted by regret theory are considerably weakened and not statistically significant. There are experimental studies that report evidence in line with regret or salience theory that do not rely on manipulations of the correlation of payoffs (e.g., Bleichrodt et al., 2010; Königsheim et al., 2019). However, these studies do not test the key behavioral assumption made in regret or salience theory in the way studies on correlation effects do. Consistent with Starmer and Sugden (1993)'s findings on regret theory, we find no evidence that correlation between acts matters as advocated by salience theory. An independent study conducted by Ostermair (2021) also suggests that ESE play a significant role in the prior empirical evidence supporting salience theory, although our results differ in certain aspects due to the distinct implementation of our experiments, as will be discussed later on. Taken together, these studies cast doubt on the descriptive ability of salience and regret theory. Notably, in a more recent investigation, Ostermair (2022) demonstrates that ESE persist under subjective uncertainty when testing the skew-symmetric property imposed by regret and salience theory within the context of the Allais paradox. Similarly, Leland et al. (2019) observe that transparent and non-transparent framings of choices can have substantial effects on risk behavior. As noted by the authors, the model of Bordalo et al. (2012) does not anticipate framing effects between minimal and transparent frames. Our study supplements their findings with additional evidence, emphasizing that controlling for decision framing is crucial when conducting experimental tests of

[5] It is plausible that regret theory and salience theory could yield disparate predictions under certain conditions. For instance, the inclusion of phantom lotteries should affect agents' choices according to salience theory, but not regret theory. We thank an anonymous referee for bringing up this point.

theories.

The remainder of this paper is organized as follows. Section 3.2 briefly discusses salience theory. In Section 3.3, we detail our experimental design. In Section 3.4, we present our results. Section 3.5 concludes.

3.2 Salience theory

In this section we provide a brief summary of salience theory (Bordalo et al., 2012). Consider a situation in which a decision maker has to choose between two lotteries, A and B. There are S states of nature denoted by $s = 1, ..., S$, each with an associated probability given by p_s. Each of the lotteries $\theta \in \{A, B\}$ assigns a payoff x_s^θ to each possible state of nature. The key feature of salience theory is a function $\sigma(x_s^A, x_s^B)$ that measures the salience of the different states. The salience function, for positive payoffs, satisfies the following conditions.[6]

Definition 2. *A salience function is a continuous and bounded function $\sigma(x_s^A, x_s^B)$ that satisfies the following conditions.*

1. *Ordering: Consider two states $s, \tilde{s}$. Denote $x_s^{min} = min\{x_s^A, x_s^B\}$ and $x_s^{max} = max\{x_s^A, x_s^B\}$. If $[x_s^{min}, x_s^{max}] \subset [x_{\tilde{s}}^{min}, x_{\tilde{s}}^{max}]$, then $\sigma(x_s^A, x_s^B) < \sigma(x_{\tilde{s}}^A, x_{\tilde{s}}^B)$.*

2. *Diminishing sensitivity: For two payoffs x_s^A and x_s^B, and any $\epsilon > 0$, $\sigma(x_s^A + \epsilon, x_s^B + \epsilon) < \sigma(x_s^A, x_s^B)$.*

As Dertwinkel-Kalt and Köster (2019) argued, the ordering property can be intuitively understood as a contrast effect. The higher the contrast between two payoffs in the same state is, the more salient the state becomes. For instance, a state with payoffs of 100 and 0, which has an outcome contrast of 100, is more salient than a state with payoffs of 10 and 0, which has a outcome contrast of 10. The property of diminishing sensitivity can be understood as a level effect (Dertwinkel-Kalt and Köster, 2019). The same contrast in payoffs will be perceived as more salient for lower levels. For instance, the difference between 0 and 10 is more salient than the difference between 100 and 110.

The salience function establishes a salience ranking $k_s \in \{1, ..., |S|\}$ which denotes the rank of state s, with lower ranks indicating greater salience. If two states receive the same salience, they obtain the same rank. There are no jumps in the ranking. State s then receives a decision weight π_s according to its ranking, i.e.,

$$\pi_s = p_s \frac{\delta^{k_s}}{\sum_{j=1}^{S} \delta^{k_j} p_j}$$

where $\delta \in (0, 1]$ is the degree of local thinking that measures to which extent salient payoffs are overweighted by the decision maker. Thus, the most salient states are overweighted, whereas the least salient states are underweighted, relative to their probabilities of occurring. Finally, the decision maker evaluates lottery A and B based on these decision weights and chooses lottery A if

[6] To allow for negative outcomes and explain the reflection effect, the salience function needs to satisfy the reflection assumption: for all $x, x', y, y' > 0$, $\sigma(x, y) < \sigma(x', y') \Leftrightarrow \sigma(-x, -y) < \sigma(-x', -y')$.

and only if

$$\sum_{s \in S} \pi_s [v(x_s^A) - v(x_s^B)] > 0$$

Dertwinkel-Kalt and Köster (2019) introduced a continuous version of salience theory and demonstrated that salience theory induces a preference for skewness, not in an absolute, but in a relative sense. As a measure for a lottery's absolute skewness, Dertwinkel-Kalt and Köster (2019) considered the third standardized central moment. The relative skewness is defined as the third centralized moment of the difference in the payoffs between lottery A and B in a given state, $\Delta_A = X^A - X^B$. Since $\Delta_B = -\Delta_A$, it follows that $S(\Delta_B) = -S(\Delta_A)$. Lottery A is said to be positively skewed relative to lottery B if $S(\Delta_A) > 0$. Intuitively, whenever a decision maker chooses between two lotteries, the salience of the payoffs of a lottery depends on the context, that is on the payoffs that co-occur in the same state.[7] A lottery with a positive relative skewness has an upside that stands out in comparison to the alternative lottery.

As preferences for absolute skewness are implied by many other theories such as prospect theory (Kahneman and Tversky, 1979), establishing that salience theory induces a preference for relative skewness greatly clarifies what distinguishes salience theory from other decision theories under uncertainty. In a controlled lab experiment, one can manipulate a lottery's relative skewness by changing its correlation while leaving its marginal distribution constant. Therefore, testing for correlation effects allows to test the core assumptions of salience theory in a very clean way.

3.3 Experimental design

In this section, we detail our experimental design, the predictions made by salience theory in the considered settings, and, where applicable, how ESE might impact choices.

3.3.1 Setting I: common consequence Allais paradox

3.3.1.1 Salience theory and the common consequence Allais paradox

Experiments that demonstrate the common consequence Allais paradox commonly invoke two choices between a relatively riskier lottery A of the form $(a_h, p_h; z, p_z; a_l, p_l)$ and a safer lottery B of the form $(b, p_h; z, p_z; b, p_l)$, where $a_h > b > a_l$.[8] In expected utility theory, the common consequence $z \in \{a_l, b\}$ is irrelevant to choices. However, a common observation is that participants in lab experiments express a preference for the safer option if $z = b$ but for the riskier option if $z = a_l$ (Allais, 1953).

Salience theory can explain this pattern if participants perceive the lotteries as independent (Bordalo et al., 2012). Moreover, salience theory implies that the Allais paradox can be turned off when payoffs are perfectly positively correlated. To see how salience theory can explain the Allais

[7] As Dertwinkel-Kalt and Köster (2019) point out, whenever one of the lotteries is a degenerate safe option, the relative skewness of the other lottery boils down to its absolute skewness.

[8] For salience theory as discussed by Bordalo et al. (2012) to yield an unambiguous salience ranking, we further impose $b - a_l > a_h - b$.

(i) Replication 1 with independent lotteries when $z = b$, and three states

	p_h	p_z	p_l
Lottery A	a_h	b	a_l
Lottery B	b	b	b

(ii) Replication 2 with positive correlation when $z = a_l$, and three states

	p_h	p_z	p_l
Lottery A	a_h	a_l	a_l
Lottery B	b	a_l	b

(iii) Replication 3 with independent lotteries when $z = a_l$, and four states

	$p_h(p_l + p_h)$	$p_h p_z$	$p_z(p_l + p_z)$	$(p_l + p_z)(p_l + p_h)$
Lottery A	a_h	a_h	a_l	a_l
Lottery B	b	a_l	a_l	b

(iv) CEO 1 with positive correlation and five states

	$p_h - p_h p_z$	$p_h p_z$	$p_z - p_h p_z$	$p_h p_z$	p_l
Lottery A	a_h	a_h	z	z	a_l
Lottery B	b	b	z	z	b

(v) CEO 2 with independent lotteries and five states

	$p_h - p_h p_z$	$p_h p_z$	$p_z - p_h p_z$	$p_h p_z$	p_l
Lottery A	a_h	a_h	z	z	a_l
Lottery B	b	z	z	b	b

Table notes: parameters satisfy the following conditions: $a_h > b > a_l$; $z \in \{a_l, b\}$; $b - a_l > a_h - b$.

Table 3.1 Experimental tasks on the Allais paradox

paradox if lottery A and B are independent, consider Table 3.1 for a visual representation of the choice under different correlation structures. Consider first the case in which $z = b$ (see Table 3.1i). Intuitively, the state (a_l, b) is the most salient one. Consequently, the low payoff of lottery A and the high payoff of lottery B are overweighted. This makes lottery A relatively unattractive when $z = b$. Now, consider the case in which $z = a_l$. When lotteries are independent (see Table 3.1iii), the state in which lottery A yields the high payoff a_h and lottery B yields the low payoff a_l is the most salient one. Consequently, the high payoff of lottery A and the low payoff of lottery B are overweighted. As a result, the riskier lottery A becomes more attractive when $z = a_l$. To understand why salient theory predicts that the Allais paradox does not occur under positive correlation, consider Table 3.1i and 3.1ii. Note that, irrespective of the value of z, whether it is b or a_l, the state (z, z) is simply disregarded by the decision-maker. Therefore, altering z cannot impact lottery choices when payoffs are perfectly positively correlated.

To see more clearly how alterations in the correlation structure impact the occurrence of the Allais paradox with a common consequence, consider the case where $z = b$. In this scenario, lottery B guarantees a payoff of b, implying a fixed correlation structure (see Table 3.1i). Consequently, any changes in the frequency of the Allais paradox, resulting from changes in the correlation structure, must arise from choices when $z = a_l$. Under maximally positive correlation, the most

salient state is when lottery A yields the low payoff a_l, and lottery B yields the medium payoff b (see Table 3.1ii). Transitioning to the case of independent lotteries, the state (a_h, a_l) becomes the most salient (see Table 3.1iii). Thus, the shift from maximally positive correlation to independence enhances the perceived attractiveness of lottery A compared to lottery B. This shift may lead to a preference change from the safer lottery B to the riskier lottery A.

3.3.1.2 ESE and the Allais paradox

In the experimental tests of correlation effects in the Allais paradox (Bordalo et al., 2012; Frydman and Mormann, 2018; Bruhin et al., 2022), choice tasks are often displayed in a state of the world representation that makes the correlation structure evident to subjects. Bruhin et al. (2022) additionally employ a different choice display that does not make the correlation structure evident when lotteries are independent. In the following, we show how ESE confound results in both types of choice displays.

In studies using the state of the world display, event-splitting tends to occur when choices between lotteries are displayed in the minimal state space. Consider again Table 3.1. Whenever payoffs are maximally positively correlated, the choice problem can be displayed in a matrix with 3 states (see Table 3.1i and 3.1ii). Note that this is always true for $z = b$. However, when $z = a_l$ and the lotteries are independent, the choice problem must be displayed in a matrix with at least four states (see Table 3.1iii). ESE occur if this difference in the number of displayed states and their probability of occurring, independent of the change in the correlation structure, impacts lottery choices.

To see how ESE might impact choices, note first that the minimal state space does not change for $z = b$. As a result, in the considered studies, no event-splitting occurs for $z = b$. Consider therefore the case in which $z = a_l$. Intuitively, changing the correlation structure from maximally positive to independent, the high payoff of lottery A (a_h) and the low payoff of lottery B (a_l) appear in one additional state. If decision makers attach more weight to payoffs that are displayed more often, irrespective of their probability, decision makers might attach more weight to the high payoff of lottery A and the low payoff of lottery B when lotteries are independent. This would render lottery A more attractive but lottery B less attractive. Thus, in the considered setting, ESE could induce similar effects as predicted by salience theory.

We now turn to the alternative choice display used by Bruhin et al. (2022). Whenever lotteries are maximally positively correlated, choices are displayed as illustrated in Table 3.1i and 3.1ii. However, when lotteries are independent, choices are in the "canonical" display as illustrated in Table 3.2. When $z = b$, three payoffs are displayed for lottery A, and lottery B is displayed as having one outcome (see Table 3.2i). When payoffs are maximally positively correlated, three payoffs are displayed for both lotteries (see Table 3.1ii). Thus, for lottery A, ESE do not occur. For lottery B, the same payoff b is displayed thrice instead of once. However, since lottery B is degenerate, ESE are unlikely to exert significant influence on choices.

Consider now the case when $z = a_l$. In the canonical display, when lotteries are independent, the high and the low payoff of lottery A and B occur once each (see Table 3.2ii). However, when

(i) For $z = b$

	p_h	p_z	p_l			100%
Lottery A	a_h	b	a_l	or	Lottery B	b

(ii) For $z = a_l$

	p_h	$p_z + p_l$			$p_h + p_l$	p_z
Lottery A	a_h	a_l	or	Lottery B	b	a_l

Table notes: parameters satisfy the following conditions: $a_h > b > a_l$; $z \in \{a_l, b\}$; $b - a_l > a_h - b$.

Table 3.2 "Canonical" display in Bruhin et al. (2022)

payoffs are positively correlated, the low payoff of the risky lottery A (a_l) is displayed twice, and the high payoff of the safer lottery B (b) is also displayed twice (see Table 3.1ii). If decision makers attach more weight to payoffs that are displayed more often, ESE could once again cause behavior similar to that predicted by salience theory.

3.3.1.3 Experimental design and hypotheses

To disentangle correlation effects from ESE, we employ two between-subjects treatments. In the replication treatment,[9] for $z = a_l$, whenever the correlation structure is changed from positive to independent, the number of states changes from three to four (i.e., Table 3.1ii and 3.1iii). For $z = b$, there are always three states. In our second treatment, the correlation effects only (CEO) treatment, subjects face the presentation formats in Table 3.1iv and 3.1v. In this treatment, there are always 5 states and the displayed probabilities of each payoff stay constant when changing the correlation structure.[10] This design controls for ESE and thus allows testing for correlation effects in a clean way. This design is inspired by Starmer and Sugden (1993) who employ treatments similar to our replication and CEO treatment to disentangle correlation effects from ESE using different choice tasks.

Each subject faces choices for 3 parameter sets (see Table 3.3) that might elicit behavior typical of the common consequence Allais paradox. For $z = a_l$, each subject faces the choice task for the two different correlation structures. For $z = b$, each subject faces only one choice. This results in $3 * (2 + 1) = 9$ decisions for each subject.

Our main hypotheses for setting I are summarized in Hypothesis 2. As discussed above, chang-

[9] It is worth noting that this part of our experiment replicates only certain aspects of the existing experiments (Bordalo et al., 2012; Frydman and Mormann, 2018; Bruhin et al., 2022). We utilize identical parameters for the lotteries, specifically for parameter set 1 (see Table 3), with numerical values adjusted to align with the stakes utilized in our other experimental tasks. While Bordalo et al. (2012) exclusively employ tasks parameterized in accordance with parameter set 1, Bruhin et al. (2022) encompass a broader spectrum of parameter values. Additionally, Frydman and Mormann (2018) encompass three distinct correlation structures - maximally positive, independent, and an intermediate correlation - in contrast to our inclusion of only the former two. It is important to mention that both Bordalo et al. (2012) and Frydman and Mormann (2018) also incorporate experimental tasks that are not specifically designed to elicit the Allais paradox.

[10] The reason for displaying choices in 5 states is to maintain a constant number of displayed states and their corresponding probabilities. To understand this, note first that the minimal state space for $z = a_l$ is given by four states with probabilities $p_h - p_z$, $p_h p_z$, $p_z - p_h p_l$, and $p_h p_z + pl$. When $z = b$, the minimal state space is given by three states with probabilities p_h, p_z, and p_l. In order to increase the number of displayed states to four, one would have to split one of the states into two. However, this would generally result in different probabilities for $z = a_l$ and $z = b$. Therefore, a minimum of five states is required to display the choice tasks without changing the number of states or the displayed probabilities when changing the correlation structure.

	p_h	p_l	p_z	a_h	a_l	b
Set 1	33%	1%	66%	125	0	120
Set 2	25%	5%	70%	95	5	76
Set 3	30%	10%	60%	145	22	104

Table 3.3 Parameter values for the common consequence Allais paradoxes

ing the correlation structure from independent to maximally positive should eliminate choice patterns in line with the Allais paradox (Bordalo et al., 2012). This motivates Hypothesis 2.1. If effects found in previous studies are indeed driven by correlation effects, we should expect this prediction to hold true also in the CEO treatment. This motivates Hypothesis 2.2. Finally, if ESE and correlation effects are additively separable, comparing choices in the two treatments allows disentangling the effects of ESE and correlation effects. This motivates Hypothesis 2.3.

Hypothesis 2.

(1) In the replication treatment, when lotteries are independent, choices will exhibit a pattern consistent with the common consequence Allais paradox more frequently than when choices are maximally positively correlated.

(2) Hypothesis 1.1 will also be confirmed in the CEO treatment.

(3) The effects found in the replication treatment are largely driven by ESE.

3.3.2 Setting II: Mao pairs

3.3.2.1 Salience theory and preferences for relative skewness

Dertwinkel-Kalt and Köster (2019) test the prediction that salience theory induces a preference for relative skewness in the context of choices between two lotteries of a Mao pair (Mao, 1970), denoted $M(E, V, S, \eta)$. Mao pairs are pairs of lotteries $L(E, V, -S)$ and $L(E, V, S)$ that share the same expected value E and variance V. Their skewnesses are of equal size but different sign (i.e., $-S$ and S). Since the first two moments are held constant, Mao pairs are particularly well suited to investigate (relative) skewness preferences. The joint distribution of the two lotteries is described by a parameter η. $\eta = 0$ corresponds to the perfectly negative correlation and $\eta = 1$ corresponds to the maximally positive correlation.

In Dertwinkel-Kalt and Köster (2019)'s second experiment, subjects decide between the six Mao pairs displayed in Table 3.4. Two of the Mao pairs each have the same variance, with one of these two Mao pairs being more symmetric (corresponding to $S = 0.6$ in Table 3.4) than the other. To each subject, each Mao pair was presented in two correlation structures, maximally positively and perfectly negatively correlated. Consider Table 3.5. Moving from the maximally positive correlation structure to the perfectly negative one, i.e., moving from Table 3.5i to Table 3.5iii, increases the relative skewness of the right-skewed lottery. Moreover, for the symmetric Mao pairs, changing the correlation structure induces a sign change in the relative skewness of the lotteries. When $\eta = 1$, it is the left-skewed lottery that is positively skewed relative to the

#	Left-skewed lottery	Right-skewed lottery	Variance	Abs. skewness	Relative skewness	
					$\eta = 0$	$\eta = 1$
1	(120, 90%; 0, 10%)	(96, 90%; 216, 10%)	1296	2.7	-2.7	-1.5
2	(135, 64%; 60, 36%)	(81, 64%; 156, 36%	1296	0.6	-0.6	1.0
3	(40, 90%; 0, 10%)	(32, 90%, 72, 10%)	144	2.7	-2.7	-1.5
4	(45, 64%; 20, 36%)	(27, 64%; 52, 36%)	144	0.6	-0.6	1.0
5	(80, 90%; 0, 10%)	(64, 90%; 144, 10%)	576	2.7	-2.7	-1.5
6	(90, 64%; 40, 36%)	(54, 64%; 104, 36%)	576	0.6	-0.6	1.0

Table 3.4 Mao pairs used in Dertwinkel-Kalt and Köster (2019)

right skewed lottery. When $\eta = 0$, this relationship is reversed. For the asymmetric Mao pairs, the right-skewed lottery is always the lottery with a positive relative skewness, regardless of the correlation structure.

Salience theory predicts that the share of subjects choosing the right skewed lottery (weakly) increases when changing the correlation of the lotteries from maximally positive to perfectly negative. Moreover, this effect is predicted to be larger for the symmetric Mao pairs. In their experiment, Dertwinkel-Kalt and Köster (2019) confirm both hypotheses.

3.3.2.2 ESE in Dertwinkel-Kalt and Köster (2019)

To see how ESE can account for these findings, consider Table 3.5. In Dertwinkel-Kalt and Köster (2019), subjects are confronted with choices analogous to Table 3.5i and 3.5iii. When moving from Table 3.5i to Table 3.5iii, the correlation is increased from $\eta = 0$ to $\eta = 1$. However, the number of displayed states also changes from two to three. The high payoff of the left skewed lottery $L(E, V, -S)$ and the low payoff of the right skewed lottery are displayed once under negative correlation but twice under positive correlation. Therefore, if multiply displayed states receive a higher decision weight, these changes in the presentation of the choice problem might induce choice patterns similar to those predicted by salience theory.

(i) $\eta = 0$

Probability	p	$1-p$
$L(E, V, -S)$	x_1	x_2
$L(E, V, S)$	y_2	y_1

(ii) $\eta = 0$

Probability	p	$1-2p$	p
$L(E, V, -S)$	x_1	x_2	x_2
$L(E, V, S)$	y_2	y_1	y_1

(iii) $\eta = 1$

Probability	p	$1-2p$	p
$L(E, V, -S)$	x_1	x_2	x_2
$L(E, V, S)$	y_1	y_1	y_2

Table note: parameters satisfy the following conditions: $y_2 > x_2 > y_1 > x_1$, $p \in (0, 1/2)$ and $\eta \in [0, 1]$.

Table 3.5 Reducing relative skewness

3.3.2.3 Experimental design and hypotheses

We again introduce two treatments. In the replication treatment,[11] whenever the correlation structure is changed from perfectly negative to maximally positive, we also introduce ESE (see Table 3.5i and 3.5iii). In the correlation effects only (CEO) treatment, we control for ESE and only correlation effects are present (see Table 3.5ii and 3.5iii). Each subject faces 6*2=12 decisions in setting II.

We test Hypothesis 3 summarized below. Hypothesis 3.1 states that in the replication treatment we expect to replicate the findings of Dertwinkel-Kalt and Köster (2019). If the effects reported in Dertwinkel-Kalt and Köster (2019) are driven by correlation effects, we should expect to find choice patterns in line with Hypothesis 3.1 also in the CEO treatment. This motivates Hypothesis 3.2. Further, comparing choices in the two treatment will allow to disentangle ESE from correlation effects. This motivates Hypothesis 3.3.

Hypothesis 3. *Consider two Mao pairs $M(E, V, S', \eta)$ and $M(E, V, S'', \eta)$ with $S' < S''$.*

(1) In the replication treatment, (a) for each of the Mao pairs the share of subjects choosing the right-skewed lottery is larger for $\eta = 0$ (i.e., the perfectly negative correlation) than for $\eta = 1$ (i.e., the maximal positive correlation). (b) The correlation effect described in (a) is larger for the more symmetric Mao pair $M(E, V, S^{'}, \eta)$.

(2) Hypothesis 2.1 will also be confirmed in the CEO treatment.

(3) The effects found in the replication treatment are largely driven by ESE.

3.3.3 Procedures

The experiment was conducted in Beijing at Renmin University of China. We preregistered our main experimental hypotheses with the AEA social science registry under the ID of AEARCTR-0007239. A total of 15 experimental sessions were conducted in March 2021, with a total of 296 Chinese undergraduate students participating.

The experiment was programmed with oTree (Chen et al., 2016) and conducted in Chinese and in a physical lab. The instructions and display of the choice tasks were modeled on Dertwinkel-Kalt and Köster (2019). The complete instructions can be found here. Payoffs were displayed in an experimental currency that was translated into Yuan at a rate of 0.5. Tasks were presented in a matrix form. See Figure 3.1 for an example. The presentation of choice tasks closely follows that of Dertwinkel-Kalt and Köster (2019). The order in which states appear, as well as which lottery was labelled option A and B, was randomized at the subject level. Participants received tasks in random order.

Our experiment used a between-subjects design. Participants in both the replication and CEO treatments decided on a total of 35 choice tasks.[12] They read that for each choice task, there were

[11] This part of our experiment follows the design of Dertwinkel-Kalt and Köster (2019)'s second experiment rather closely. It is worth noting that Dertwinkel-Kalt and Köster (2019) is mainly a theory paper that also incorporates supporting experiments.

[12] Except the choice tasks described in this paper, we also included another 14 choice tasks for a companion study.

Please choose between options A and B.

	Fields 1-36 (36.0%)	Fields 37-64 (28.0%)	Fields 65-100 (36.0%)
○ Option A	135	135	60
○ Option B	156	81	81

Figure 3.1 Example of a decision screen

Figure notes: "Fields" refers to the fields of a wheel of fortune.

two options with payoffs that depend on the turn of a wheel of fortune.[13] One of these tasks was randomly selected and paid out at the end of the experiment. In addition, participants received a show-up fee of 10 Yuan. Participants took about 30 minutes to complete the experiment and received an average payment of around 41 Yuan.

3.4 Results

In this section, we present the results of the experiment. We start with the setting of the Allais paradox and then move on to the setting of Mao pairs. For both settings, our findings suggest that existing evidence for salience theory is mainly driven by ESE.

3.4.1 Setting I: common consequence Allais Paradox

Since we find that the occurrence of the common consequence Allais paradox varies significantly between our three parameter sets, we present results for each lottery separately.[14] We begin with the replication treatment. Figure 3.2a displays the occurrence of the Allais paradox choice pattern, net of the reverse choice pattern, for the three parameter sets for both the positive correlation structure and the case of independence. For all three parameter sets, we find that changing the correlation structure from positive to independent leads to a significant ($p < 0.01$ for each choice task, two-sided t-test[15]) and large increase in choice patterns consistent with the Allais paradox. These patterns are generally in line results reported in recent studies (Bordalo et al., 2012; Frydman and Mormann, 2018; Bruhin et al., 2022).

To investigate the role of pure correlation effects, we next turn to the frequency of Allais-consistent choice pattern in the CEO treatment. See Figure 3.2b. For parameter set 1, changing the correlation of the lotteries from positive to independent increases the occurrence of the Allais paradox significantly ($p = 0.049$) from about 16% to 27%. This finding is in line with the predictions of salience theory. For parameter set 2, we observe that 11% (9.5%) of subjects exhibit a choice

[13] For the Allais paradox choices, we round probabilities to multiples of 1%. We do so in a way that induces a more negative correlation than independence would imply. If anything, this rounding should make it more likely for choice patterns as predicted by salience theory to arise.

[14] At this point, we deviate from our pre-analysis plan which had specified only an analysis that pools all the three parameter sets.

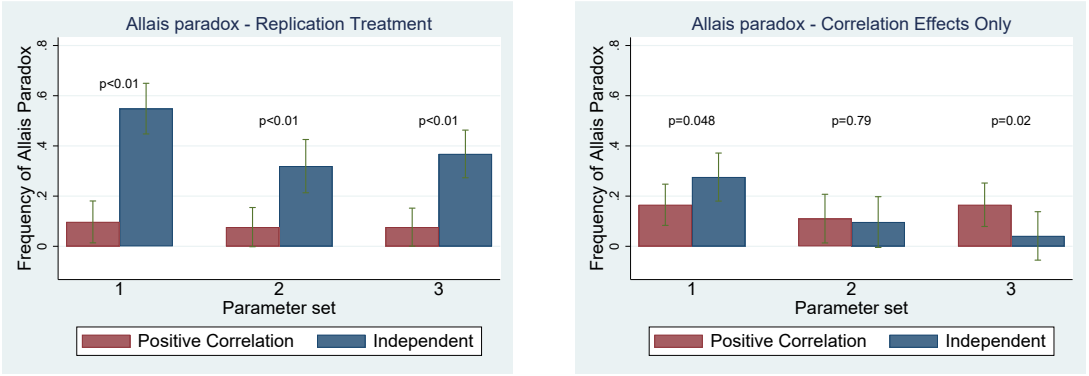
[15] Unless otherwise noted, the p-values stated in this paper result from two-sided t-tests. Whenever we have repeated observations from the same individual, we cluster at the subject level.

pattern in line with the Allais paradox when lotteries are positively correlated (independent). This difference is not statistically significant ($p = 0.79$). Finally, for parameter set 3, changing the correlation structure from positive to independent reduces the occurrence of the Allais paradox significantly ($p = 0.012$) from about 16% to 3%. This effect goes in the opposite direction as predicted by salience theory.

Comparing Figure 3.2a and 3.2b, it is evident that the reactions to changes in the correlation structure are much more pronounced but also more systematic in the replication treatment than in the CEO treatment. This indicates that in our experiment, ESE are a much more important driver of behavior than correlation effects.

Now we test Hypothesis 2 more formally. We define a variable $shift$ as $R(a_l \mid independent) - R(b \mid independent) - \big(R(a_l \mid positive) - R(b \mid positive)\big)$, where $R(z \mid cs)$ is a dummy that equals 1 if a subject chose the risky option for common consequence $z \in \{a_l, b\}$ under correlation structure $cs \in \{independent, positive\}$, and 0 zero otherwise. For a given correlation structure, $R(a_l \mid cs) - R(b \mid cs) = 1$ indicates the classical Allais choice pattern, whereas a value of -1 indicates the reverse choice pattern and a value of 0 indicates behavior consistent with expected utility theory. Since the correlation structure does not change when $z = b$, i.e., $R(b \mid independent) = R(b \mid positive)$, the expression of $shift$ boils down to $R(a_l \mid independent) - R(a_l \mid positive)$. That is, changes in the frequency of the occurrence of the Allais paradox due to changes in the correlation structure must therefore stem from correlation effects in the case in which $z = a_l$. On the aggregate, the variable $shift$ measures the increase in Allais paradox compatible behavior as a result of changing the correlation structure as predicted by salience theory, net of the reverse choice pattern. Notice that by netting out behavior consistent with the reverse Allais choice pattern, we are able to control for choice reversals that stem from decision noise.

We run a linear regression of the variable $shift$ on a constant and a dummy that is equal to one if a given subject was in the replication treatment.[16] The constant provides an estimate of the change in the frequency of the Allais paradox that is due to correlation effects. The coefficient



(a) Replication Treatment **(b)** CEO Treatment

Figure 3.2 Frequency of the Allais paradox (net of the reverse choice pattern)

Figure notes: there are 149 subjects in the replication treatment and 147 subjects in the CEO treatment. Values of parameter sets can be found in Table 3.3.

[16] The linear regression model yields t-tests for the variable $shift$ in the different treatments. Clustering at the individual level allows to account for repeated observations when applicable.

on the replication dummy provides an estimate of the additional ESE. Since we find different choice patterns for the different parameter sets, we run this regression for all three parameter sets separately, as well as jointly.

We report our regression results in Table 3.6. We find that for the first parameter set, changing the correlation from positive to independent (controlling for ESE) induces an increase in the frequency of Allais compatible behavior by about 11 percentage points. The coefficient is significant at the 5% level. ESE induce a further increase in the frequency of Allais consistent behavior by around 35 percentage points ($p < 0.01$). For the second parameter set, we find no evidence that the frequency of Allais compatible behavior is influenced by changing the correlation of the lotteries from positive to independent. The point estimate of the constant is -0.01 ($p = 0.79$). The coefficient on the replication dummy is estimated at 0.255 ($p < 0.01$), which implies that ESE lead to an increase in the frequency of the Allais paradox of around 26 percentage points. Finally, for the third parameter set, we find that changing the correlation of the lotteries from positive to independent (controlling for ESE) leads to a decrease in the occurrence of Allais compatible choice patterns by about 14 percentage points ($p = 0.01$). We estimate that ESE lead to an increase in the frequency of Allais compatible behavior by around 42 percentage points. Finally, pooling the data from all three parameter sets, we find an estimate of the average correlation effect for our three decision tasks that is close to zero and not statistically different from zero ($p = 0.64$). On average, ESE induce a large and significant increase in the Allais paradox by 34 percentage points ($p < 0.01$). We summarize our results as follows:

Result 2. *On the Allais paradox:*

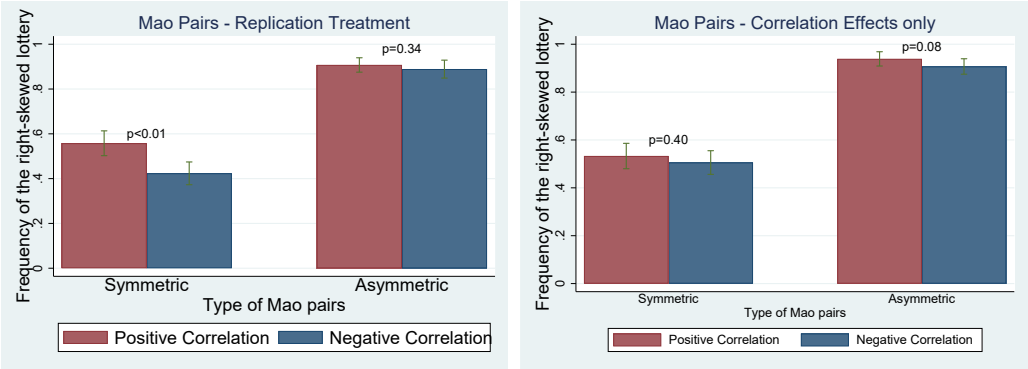
(1) In the replication treatment, changing the correlation structure from maximally positive to independent increases the frequency of the common consequence Allais paradox for all three parameter sets. We confirm Hypothesis 2.1.

(2) In the CEO treatment, we observe the above pattern for parameter set 1. For parameter set 2, no evidence of correlation effects is found. For parameter set 3, we find that changing the correlation from maximally positive to independent reduces the frequency of the common

	(1) Parameter set 1	(2) Parameter set 2	(3) Parameter set 3	(4) Pooled
Replication	0.348***	0.255***	0.418***	0.340***
	(0.074)	(0.075)	(0.074)	(0.046)
Constant	0.109**	-0.014	-0.136**	-0.0136
	(0.055)	(0.051)	(0.053)	(0.029)
Observations	296	296	296	888
Individuals	296	296	296	296
R-squared	0.071	0.038	0.098	0.066

Table notes: robust standard errors are in parentheses. For regression (4), standard errors are clustered at the subject level. Notations for significance levels are as follows: * for p≤0.1; ** for p≤0.05; *** for p≤0.01. The values for parameter sets 1-3 can be found in Table 3.3.

Table 3.6 Relative impacts of ESE and correlation effects in the setting of the Allais paradox



(a) Replication Treatment **(b)** CEO Treatment

Figure 3.3 Frequency of choices of the right-skewed option

Figure notes: there are 149 subjects in the replication treatment and 147 subjects in the CEO treatment. For each type of Mao pair, there are three Mao pairs. The parameters for the Mao pairs can be found in Table 3.4.

consequence Allais paradox. As we do not find consistent evidence for correlation effects as predicted by salience theory, we reject Hypothesis 2.2.

(3) The effects in the replication treatment are largely driven by ESE. This result confirms Hypothesis 2.3.

3.4.2 Setting II: Mao pairs

We begin the exposition of our findings with the replication treatment, where both ESE and correlation effects are introduced at the same time. For the asymmetric Mao pairs,[17] 90% of choices were for the right skewed option when the correlation was maximally positive. When the correlation was changed to negative, this hardly impacted choices. See Figure 3.3. These findings are very much in line with the results of Dertwinkel-Kalt and Köster (2019), both with respect to the choice frequencies and the lack of reaction to changes in the correlation structure. For the more symmetric Mao pairs, changing the correlation structure from positive to negative decreases the choice frequency of the positively skewed option from 56% to 42%. This difference is significant at the 1% level. However, the effect goes in the opposite direction as predicted by salience theory and as the effect reported by Dertwinkel-Kalt and Köster (2019).

We next turn to the CEO treatment, in which ESE are well controlled for. For the asymmetric Mao pairs, 94% of choices are for the positively skewed option when the correlation is positive. When lotteries are negatively correlated, 91% of choices are for the positively skewed lottery. This difference is marginally statistically significant ($p = 0.08$) and goes in the opposite direction as predicted by salience theory. Turning to the more symmetric Mao pairs, 53% of choices were for the positively skewed option when the correlation was positive. This number is equal to 51% when the correlation is changed to negative. This difference is not statistically significant at any conventional significance level ($p = 0.38$). When ESE are controlled for, we find no evidence that changing the correlation structure induces systematic and meaningful changes in lottery choices.

[17] Since choice patterns do not substantially differ for the different Mao pairs, we pool all symmetric and asymmetric Mao pairs for this part of the analysis.

	(1) Replication	(2) Dertwinkel-Kalt and Köster (2019)	(3) CEO	(4) Pooled
asymmetric	0.125***	-0.118***	-0.005	-0.005
	(0.042)	(0.038)	(0.035)	(0.035)
Replication				-0.114**
				(0.050)
Replication $*$ *asymmetric*				0.130**
				(0.054)
Constant	-0.141***	0.127***	-0.028	-0.028
	(0.039)	(0.034)	(0.032)	(0.032)
Observations	894	1,152	882	1,776
R-squared	0.013	NA	0.000	0.009
Individuals	149	192	147	296

Table notes: standard errors clustered at a subject level are in parentheses. Notations for significance levels are as follows: * for p≤0.1; ** for p≤0.05; *** for p≤0.01.

Table 3.7 Relative impacts of ESE and correlation effects in the setting of Mao pairs

Comparing the choice patterns in the replication and the CEO treatment suggest that the effects we observe in the replication treatment are mainly driven by event-splitting, and not changes in the correlation structure. To test Hypothesis 3 more formally, we follow the analysis of Dertwinkel-Kalt and Köster (2019) closely. We construct a variable $shift$ that is equal to one if a given subject chose the left skewed option for the maximally positive correlation but shifted to the right skewed option under negative correlation. For the reverse choice pattern, $shift = -1$. Finally, $shift = 0$ when a subject chose the same lottery for both correlation structures. For each treatment, we then regress $shift$ on a constant and dummy that is equal to 1 if a given Mao pair is asymmetric, and 0 otherwise.

Table 3.7 summarizes our regression results. In column 1, we report results for the replication treatment. The constant is estimated to be -0.14 and is statistically significant at the 1% level. In line with the exposition above, this suggests that changing the correlation structure from positive to negative induces subjects to switch away from the right skewed option. The estimate for the coefficient of the dummy indicating the asymmetric Mao pairs is 0.125 and is statistically significant at the 1% level. This implies that the effect of changing the correlation structure is negated for the asymmetric Mao pairs. In column 2, we report the estimates reported by Dertwinkel-Kalt and Köster (2019) for reference.Contrary to our findings, Dertwinkel-Kalt and Köster (2019) reported a positive constant of 0.127 and a negative coefficient for the dummy indicating the asymmetric Mao pairs of -0.118. As we do not observe the same choice patterns in our data as in Dertwinkel-Kalt and Köster (2019), we reject Hypothesis 3.1. Since our experimental design followed that of Dertwinkel-Kalt and Köster (2019) closely, this result is rather surprising. In column 3, we report the estimates for the CEO treatment. Here, the estimates for the constant and the dummy are both not significantly different from zero. As we find no evidence for correlation effects once ESE are controlled for, our results confirm Hypothesis 3.2.

Finally, in column 4, we pool the data of both treatments. We regress the variable $shift$ on a constant, a dummy indicating an asymmetric Mao pair, a dummy indicating the replication treatment, and an interaction between the two dummy variables. The constant and the dummy

indicating an asymmetric Mao pair provide an estimate of the frequency of preference reversals due to a change in the correlation structure. Both these coefficient are not significantly different from zero, which confirms our previous analysis. The replication dummy provides an estimate of the preference reversals that are due to ESE, net of correlation effects. This coefficient is estimated at -0.11 and is statistically significant at the 5% level. The coefficient of the interaction term is estimated at 0.13 and is statistically significant at the 5% level, which negates the ESE for the asymmetric Mao pairs. These results suggest that in our setting preference reversals are mainly driven by ESE and not by correlation effects. We therefore conclude that our results provide evidence in favor of Hypothesis 3.3. We summarize our results as follows:

Result 3. *On Mao Pairs:*

(1) In the replication treatment, participants chose the right skewed lottery more often under the maximally positive than under the maximally negative correlation structure. This effect was observed only for the symmetric Mao pairs but not for the asymmetric ones. This contradicts the findings reported by Dertwinkel-Kalt and Köster (2019). We reject Hypothesis 3.1.

(2) Once ESE are controlled for, we do not find evidence of correlation effects. We reject Hypothesis 3.2.

(3) The effects found in the replication treatment seem to be largely driven by ESE. We confirm Hypothesis 3.3.

Our failure to replicate the results of Dertwinkel-Kalt and Köster (2019) is somewhat surprising. Although our experimental design was close to that of the original paper, there are notable differences, such as the subject pool (Chinese versus German students),[18] the display of choice tasks,[19] the number of choice tasks, the incentive structure,[20] and how uncertainty was resolved.[21] However, none of these differences seems able to explain the differences between our findings. In this context, it is worth mentioning that by now three sets of authors, Dertwinkel-Kalt and Köster (2021), Ostermair (2021), and ourselves, have implemented their version of our experiment on ESE for the Mao pairs.[22] All three implementations yielded somewhat different results, despite all studies being reasonably powered. Ostermair (2021) and Dertwinkel-Kalt and Köster (2021) replicate the original findings of Dertwinkel-Kalt and Köster (2019), whereas we do not. The choice patterns reported by Dertwinkel-Kalt and Köster (2021) for their and our respective CEO treatment are

[18] Differences in risk attitudes between Chinese and Western subjects have been documented before (Hsee and Weber, 1999; Bruhin et al., 2010a). However, the findings of these studies cannot explain the differences in behavior between our study and that of Dertwinkel-Kalt and Köster (2019).

[19] Whereas we randomized the order in which states appeared on participants' screens, Dertwinkel-Kalt and Köster (2019) used a fixed order. Further, along with the fields of the wheel of fortune that determined which state of the world would materialize, we also displayed the probability of each state, whereas Dertwinkel-Kalt and Köster (2019) displayed only the numbers of the fields corresponding to different states of the world, but not the probabilities. We decided to display the probabilities of states because this makes it easier for subjects to comprehend the tasks.

[20] Subjects in our experiment decided on a total of 35 choice tasks whereas the experiment in Dertwinkel-Kalt and Köster (2019) contained only 12 choice tasks. Both studies relied on the random incentive mechanism. The final earnings relative to life expenses were somewhat higher in our study. However, since we included more choice tasks than Dertwinkel-Kalt and Köster (2019), it is not straightforward to compare the strength of incentives.

[21] To control for correlation effects due to other-regarding preferences, we emphasized that the resolution of uncertainty would be done for each subject independently. Dertwinkel-Kalt and Köster (2019) did not provide information on how uncertainty was resolved.

[22] Dertwinkel-Kalt and Köster (2021) conducted their experiment after an exchange with us, whereas Ostermair (2021) conducted his experiments independently of us.

qualitatively similar in that no consistent evidence for correlation effects is found, whereas Ostermair (2021) documents significant correlation effects contradicting regret and salience theory. Taken together, these results suggest that behavior in the considered setting might be either noisy or driven by subtle differences in design. Importantly, none of the three studies finds evidence for correlation effects as predicted by salience theory once ESE are controlled for.

3.5 Conclusion

In this paper, we report on an experiment conducted to study correlation effects in risk taking in both the setting of the Allais paradox and of Mao pairs, while controlling for ESE. In the setting of the Allais paradox, our results indicate that alterations in the choice display, rather than shifts in the correlation structure, may be responsible for the outcomes observed in prior studies by Bordalo et al. (2012), Frydman and Mormann (2018), and Bruhin et al. (2022). In the setting of Mao pairs, the picture is less clear since the results of our replication treatment contradict the results of (Dertwinkel-Kalt and Köster, 2019). However, our failure to replicate their results, together with our null findings on correlation effects once controlling for ESE, do shed considerable doubt on the experimental results of Dertwinkel-Kalt and Köster (2019). Overall, our study questions the validity of salience theory in describing risky behavior. Finally, given its substantial impact on risk behavior, researchers should be well aware of it when designing experiments involving event-splitting. Further research might be needed to better understand the mechanisms behind ESE. In the setting considered here and in other studies (Starmer and Sugden, 1993; Ostermair, 2021, 2022), ESE could be driven both by the change in the displayed number of states as well as the change in the displayed probabilities.[23] Future research could aim at disentangling these mechanisms.

Acknowledgements We would like to thank Mats Köster for fruitful discussions. We thank Astrid Hopfensitz and attendants of the BID workshop at TSE, the environment and behavior workshop at RUC and the CCBEF seminar at South Western University of Finance and Economics for helpful comments and suggestions. We thank Wanxing Dong for her excellent research assistance. The project leading to this publication has received funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University-A*MIDEX.

Declarations The authors have no conflicts of interest to declare.

[23] We are thankful to an anonymous referee for pointing this out.

Chapter 4

Outcome bias and delegated decision-making: Theory and Experiment

Moritz Loewenfeld[1][2]

Abstract

In a theory-guided experiment, I study how outcome bias, a tendency of principals to reward and punish economic agents as if they could have anticipated a random state of the world, shapes the incentives and choices of agents. Agents choose between two lotteries on behalf of their principal. One lottery is first-order stochastically dominant, but the dominated lottery is more likely to yield a higher payoff state-by-state. Despite perfectly observing the agent's choice, principals tend to reward agents if they choose the lottery, which realizes a higher payoff. As a result, they incentivize agents to choose the dominated lottery. Although most agents anticipate these incentives, only strategically sophisticated agents tend to choose the dominated lottery when they believe they have an incentive to do so. Structural estimation suggests that principals are either fully outcome-biased or fully unbiased, with less cognitively sophisticated principals displaying more outcome bias. The results imply that outcome bias might be most relevant in settings where sophisticated agents meet relatively unsophisticated principals.

Keywords: Outcome Bias; Choice under risk; Correlation sensitivity; Principal-Agent; Reciprocity; Experiment;

JEL Classifications: C91; D81; D91

[1] Toulouse School of Economics, University Toulouse Capitole, 1 Esplanade de l'Université, Toulouse 31080 Cedex 06, France. Email(✉): moritz loewenfeld@tse-fr.eu.

[2] I thank Astrid Hopfensitz and Sébastian Pouget for invaluable guidance on this project. I thank (in random order) Jiakun Zheng, Ingela Alger, Pau Juan Bartroli, Esteban Munoz Sobrado, Sophie Moinas, Christophe Bisière, Jean Tirole, Andrew Rhodes, Gabriele Camera, David Rojo Arjona, Jared Rubin, Jacques Crémer, Aurélien Baillon, Xavier Gabaix, Maximilian Müller, Pascal Lavergne, and Roland Bénabou, for fruitful discussions. I also thank attendants of the BID and BEE workshop at TSE.

4.1 Introduction

In a range of settings, decision-makers, such as politicians, CEOs, or investment managers, appear to be rewarded and punished for outcomes that are beyond their control.[3] For instance, politicians are held accountable for natural disasters or shark attacks (Achen and Bartels, 2017). Meanwhile, lab experiments document large and persistent outcome bias (OB) in ex-post evaluations. Decision-makers tend to be judged as if they could have foreseen a random outcome, even when choices are perfectly observable and the outcome is determined by a coin flip (Baron and Hershey, 1988; Gurdal et al., 2013). Rewarding decision-makers based on luck rather than their actions could be costly if it induces decision-makers like politicians or CEOs to make suboptimal choices. To provide one figure, Healy and Malhotra (2009) estimate that insufficient government investment in natural disaster preparedness costs the US taxpayer around $4.07 billion in a typical year. It is thus crucial to understand whether and how decision-makers anticipate and respond to outcome-biased rewards and punishments.

In this paper, I employ a theory-guided lab experiment to study whether outcome-biased rewards can induce decision-makers to take suboptimal actions. The following example illustrates the main ideas. A politician must decide whether to expend public funds to build a dike. In the unlikely event of a flood, the dike will protect citizens' properties and lives. Consider an outcome-biased voter who judges the politician as if they could have foreseen whether a flood will occur. If no flood occurs, building the dike seems like a waste of money for the outcome-biased voter. Consequently, they re-elect the politician only if they did not build the dike. Following a similar logic, if a flood has occurred, the voter re-elects the politician only if they did build the dike. Since a flood is unlikely to occur, on average the voter might be more likely to re-elect the politician if they do not build the dike. This can hold true even if the voter has an ex-ante preference for building the dike. If the politician anticipates such outcome-biased voting, they might find it in their best interest not to build the dike.

I consider a stylized setting of delegated decision-making that captures key elements of situations with incomplete or non-existent ex-ante contracts. I abstract away from classical principal-agent problems, such as information asymmetries, to focus on distortions that arise due to OB. The agent ("he") chooses between a first-order-stochastic-dominant (FOSD) and a dominated lottery on behalf of the principal ("she"). The principal observes the choice and the realized state of the world. The obtained payoff accrues exclusively to the principal. The principal then decides whether to allocate a fixed bonus to the agent or not. Interactions are one-shot. There is no ex-ante contract, which forces agents to form expectations about the principal's bonus decisions when contemplating their lottery choices.

I propose a simple model of OB. The principal is motivated by reciprocity to reward the agent if he makes a choice that aligns with her preferences over the lotteries. To determine whether the agent chose her preferred lottery, the principal computes the expected utilities of the two lotteries

[3] This has been documented in a variety of settings, ranging from the corporate world (Bertrand and Mullainathan, 2001; Jenter and Kanaan, 2015), politics (Healy et al., 2010; Healy and Malhotra, 2013; Achen and Bartels, 2017) for a review), professional sports (Lefgren et al., 2015; Gauriot and Page, 2019), finance (Heuer et al., 2017).

but does so in an outcome-biased way. In line with findings from the psychology literature (Baron and Hershey, 1988), I assume that the principal overweights the realized outcomes relative to their objective probability of occurring. This induces the principal to evaluate the agent as if he could have known the random outcome. In the extreme case of full outcome bias, the principal simply awards the bonus whenever the agent chooses the lottery that realizes a higher outcome. This way of rewarding can induce perverse incentives. Whenever the principal's least preferred lottery is more likely to yield a higher payoff than her preferred one, a sufficiently high level of OB leads the principal to incentivize the agent to choose her least preferred lottery. This can be the case even when one lottery first-order stochastically dominates the other.

I study this setting in a controlled lab experiment. In a first step, I test whether OB in the principals' reward decisions can indeed induce perverse incentives, that is, incentives for decision-makers to choose a suboptimal action. Second, I study to which extent agents anticipate the principals' OB. Third, I study agents' choices and link them to their beliefs.

The experimental design follows the setting of the model closely. Participants are randomly assigned to the role of principal or agent. Agents choose between pairs of dominant and dominated lotteries on behalf of their matched principal. Principals observe the agent's choice and the realized state of the world. They then decide whether to award a fixed bonus to the matched or a randomly chosen agent in the session (Gurdal et al., 2013). Agents' beliefs about the principals' bonus decisions are elicited in an incentive-compatible way (Hossain and Okui, 2013; Danz et al., 2022).

This design has several advantages. The use of FOSD lotteries implies that, regardless of risk preferences, there is one clearly optimal option from the principal's point of view. Binarized bonus decisions and the one-shot nature of the interaction imply that the optimal choice of a self-interested agent boils down to choosing the lottery that provides the higher expected probability of obtaining the reward. This allows for a clear interpretation of the agent's monetary incentives.

	1(1/3)	2(1/3)	3(1/3)
G	11	6	1
B	0	10	5

(i) Correlation Structure 1

	1(1/3)	2(1/3)	3(1/3)
G'	11	6	1
B'	5	0	10

(ii) Correlation Structure 2

Table notes: The first row denotes the different states of the world and their respective probabilities of occurring. The second and third rows denote the payoffs of the lotteries G and B.

Table 4.1 Two example choices between a first-order-stochastic-dominant and a dominated lottery.

The identification of OB and anticipated OB relies on a within-subject manipulation of how payoffs are correlated across states. Consider the example task in Table 4.1. Lottery $G(ood)$ dominates lottery $B(ad)$. Under both correlation structures, the marginal distribution of the lotteries is the same. However, under correlation structure 1, the dominant lottery yields a higher payoff in only 1/3 of the states, whereas it yields a higher payoff in 2/3 of the states under correlation structure 2. A sufficiently outcome-biased principal is thus more likely to award the bonus if an agent chooses the dominated lottery under correlation 1 but not under correlation 2. I employ four different lottery pairs analogous to that displayed in table 4.1 that provide sufficient

identifying variation to estimate the degree of outcome bias structurally.

The change in correlation structure allows disentangling the OB from other motives that could render bonus decisions outcome-dependent. As the obtained payoff remains constant across correlation structures, any change in the principals' behavior cannot be caused by distributional preferences (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), a risk-sharing motive (Charness et al., 2004), or outcome-based reciprocity (Falk and Fischbacher, 2006).

In order to separate OB and anticipated OB from any other effects that could arise from the change in the correlation structure, I implement two between-subject treatments, the reward-after and the reward-before treatment. The change in the correlation structure is implemented in both treatments. The only difference is that principals in the reward-after treatment can condition their bonus decisions on choices and outcomes, whereas their counterparts in the reward-before treatment can condition them on choices only. The reward-before treatment provides a baseline of behavior free of outcome bias. Comparing how the change in the correlation structure impacts the principals' bonus decisions and the agents' beliefs and lottery choices across the two treatments allows disentangling OB from other effects arising from the change in the correlation structure. In addition to the between-subject treatments, I also elicit principals' preferences between the lotteries employed in the principal-agent interaction at the end of the experiment. This serves as a control for the contingency that ex-ante preferences are influenced by the change in the correlation structure.

Results: I confirm the prediction that OB can cause perverse incentives. I find that the ex-post outcome comparison strongly influences principals' bonus decisions in the reward-after treatment, thus confirming results of Gurdal et al. (2013). As a result, the average principal is more likely to reward the agent for choosing the dominated lottery when it yields a higher outcome more often. At the individual level, 71% of the principals who provide strict incentives to choose either lottery are more likely to reward a dominated lottery choice, even though they choose the dominant lottery for themselves at the end of the experiment. When the dominant lottery is more likely to yield a higher payoff, I do not observe this pattern but revealed preferences and incentives are well aligned. Perverse incentives are also absent in the reward-before treatment. Instead, principals provide strong incentives to choose the dominant lottery regardless of the correlation structure.

The agents' stated beliefs suggest that they anticipate the OB in the principals' bonus decisions and the resulting incentives. A majority of the agents in the reward-after treatment expects to be strictly more likely to receive the bonus when choosing the dominated lottery. Crucially, this is the case only when it is more likely to yield a higher payoff. When it is the dominant lottery that yields a higher payoff most of the time, a large majority believes to have positive incentives to choose the dominant lottery. In the reward-before treatment, a similarly high fraction of agents anticipates positive incentives to choose the dominant lottery under both correlation structures.

Although a majority of the agents anticipates incentives to choose the dominated lottery, only strategically sophisticated agents appear to act on these beliefs. This finding suggests that strategic sophistication might be a necessary condition for anticipated OB to translate into actions that are detrimental to the principals. Despite pronounced differences in beliefs across treatments, lottery

choices are remarkably similar at the aggregate level. Even among those who believe they have strict incentives to choose the dominated lottery, a majority chooses the dominant one.[4] I provide evidence that the inconsistency between beliefs and choices is at least partially driven by a failure of strategic reasoning when agents make choices, similar to what was observed by Costa-Gomes and Weizsäcker (2008) when studying stated beliefs and actions in normal-form games. Consistent with this interpretation, I find that agents who score high in cognitive reflection (Frederick, 2005; Toplak et al., 2014) and Economics majors, all of whom have had some exposure to game theory at the time of the experiment, are significantly more likely to choose the dominated lottery when they believe they have incentives to do so.

To compare the anticipated degree of OB with the principals' actual degree of OB, I estimate the OB model structurally. If anything, my results suggest that agents tend to slightly overestimate the principals' outcome bias, although the differences in the estimates are only marginally statistically significant. I find that the representative principal bases bonus decisions 59% on outcomes and 41% on choices, whereas the representative agent anticipates a representative principal who bases bonus decisions 79% on outcomes.

Individual-level estimates reveal stark heterogeneity in OB. Moreover, a link between OB and cognitive reflection suggests that OB might be particularly prevalent in settings in which principals are relatively unsophisticated. Individual-level estimates and finite mixture models suggest that most principals are either fully outcome-biased or fully unbiased, with roughly 66% being fully outcome-biased. High individual-level estimates of the degree of OB are strongly positively associated with low levels of cognitive reflection and fast response time. This findings can also be viewed as corroborating the view that OB in bonus decisions is, in essence, the result of a "mistake" and not an expression of the principals' true preferences.

4.1.1 Relation to the literature

The results suggest that anticipated OB might induce strategically sophisticated agents to choose welfare-decreasing strategies when these tend to outperform alternative strategies. My paper links to a large literature documenting phenomena consistent with the OB in a range of settings, many of which presumably involve sophisticated agents. While rewarding and punishing lucky outcomes is often presumed to be harmful, not much is known about its impact on agents' behavior. My results suggest that politicians might be tempted to spend too little on disaster preparedness (Healy and Malhotra, 2009). While it has been argued that managers often take too little risk (Koller et al., 2012; Lovallo et al., 2020), anticipated OB might induce them to take on too little risk if success is unlikely but highly rewarding but too much risk in the opposite case. This provides an alternative explanation for the reluctance to engage in high-risk high-return R&D (Krieger et al., 2022). Similarly, OB offers a new perspective on why financial agents often take severely negatively skewed risks. Moreover, if outcome-biased investors tend to purchase funds that outperform the

[4] Models of distributional preferences do not convincingly rationalize this behavior. For altruism (Becker, 1976) to rationalize the agents' behavior, many of the agents would have to assign more weight to the principals' payoff than to their own. Inequity aversion (Fehr and Schmidt, 1999) implies that agents would prefer the principal to obtain the dominated lottery because this reduces disadvantageous inequality.

market most of the time, this provides a rationalization for why assets with a positive correlation to market returns appear to be over-prized (Karceski, 2002; Frazzini and Pedersen, 2014).

My paper contributes to a growing experimental literature that studies how OB shapes ex-post rewards and punishments in principal-agent settings but has not yet focused on how the OB shapes agents' incentives and choices (Gurdal et al., 2013; de Oliveira et al., 2017; Brownback and Kuhn, 2019; Aimone and Pan, 2020; König-Kersting et al., 2021). Most closely related to the present paper, Gurdal et al. (2013) find that ex-post reward decisions in a setting of delegated risk-taking depend on a counterfactual comparison between the obtained and the forgone outcome, even when principals can observe choices perfectly. I advance this literature in several ways. First, I provide a simple framework and show that OB can produce perverse incentives. Second, using structural estimation of the proposed model, I am the first to quantify the OB and characterize heterogeneity in OB. Third, I provide evidence that links OB to cognitive channels. Fourth, I study in detail to which extent agents anticipate the OB and act on their beliefs, which is important to understand welfare implications.

My results help reconcile findings that reciprocity does not tend to be outcome- but primarily intention-based (Charness and Levine, 2007; Davis et al., 2017; Gago, 2021; Friedrichsen et al., 2022; Chan and Wolk, 2023)[5] with the pervasive OB in reward decisions observed in other strands of literature. In the settings considered in the reciprocity literature, one of the agent's actions always yields a (weakly) better outcome for the principal than the alternative. In my model, this implies that the ex-post outcome comparison is always aligned with the ex-ante comparison based on choices, which limits the scope of OB.

My results also suggest an interesting twist to the discussion on whether selection might drive behavioral biases out of markets (Russell and Thaler, 1985; Fehr and Tyran, 2005). It is typically argued that selection of sophisticated individuals into positions of high importance could reduce the effect of behavioral biases in the aggregate (Fehr and Tyran, 2005; Enke et al., 2023). Contrary to this argument, my results suggest that the selection of sophisticated individuals into the role of economic agent might make the OB appear more prominently in the aggregate because these individuals are more likely to best-respond to their principals' OB.

Finally, the paper borrows ideas and relates to the literature on correlation-sensitivity in individual risk-taking (Loomes and Sugden, 1982; Bordalo et al., 2012; Lanzani, 2022; Loewenfeld and Zheng, 2023). It further relates to the literature on delegated risk-taking (Pollmann et al., 2014; Kling et al., 2023), research that focuses on behavioral aspects of principal-agent interactions (Eliaz and Spiegler, 2006; Fehr et al., 2007; Kőszegi, 2014), and research on hindsight bias and information projection (Camerer et al., 1989; Madarász, 2012; Danz et al., 2015; Danz, 2020).

The remainder of this paper is organized as follows. Section 2 introduces the model. The experimental design and hypotheses are introduced in section 3. Section 4 details the experimental procedures. Non-parametric results are presented in section 5. The structural estimation is

[5] An exception is Rubin and Sheremeta (2016), who find that outcome-based reciprocity renders principals' bonus decisions more random, which induces agents with a convex cost function to lower their effort. However, Davis et al. (2017) fail to replicate these results and argue that they might be due to random differences in behavior between treatments early in the experiment that lead to a hysteresis effect.

discussed in section 6. Section 7 concludes.

4.2 Setting and model

In this section, I first introduce the experimental setting. The setting deliberately abstracts away from many aspects usually present in principal-agent settings, such as asymmetric information, or the reasons for delegating decisions to the agent, so as to study the principals' OB and the agents' anticipation thereof in isolation. I then present the OB model, which helps organize thinking about behavior in the setting, and enables structural estimation of the principals' degree of OB as well as the agent's beliefs thereof.

4.2.1 Setting

An agent ("he", indexed $a \in \{1, ..., A\}$) has to choose one of two lotteries. I adapt the description of risk as defined by states of nature à la Savage (1954) that is common in Regret and Salience Theory (Loomes and Sugden, 1982; Bordalo et al., 2012). There is a finite number of S states of nature denoted by $s = 1, ..., S$, each occurring with an objective probability q_s. There is a finite set of outcomes X. A lottery $\theta \in \Theta$ is defined as a function that assigns a real-valued payoff $x_s^\theta \in X$ to each possible state of nature. In most of what follows, I consider a situation in which the agent has to choose between two lotteries denoted by G and B. See Table 4.2 for a visualization of the state-space representation of risk.

	$s = 1$ (q_1)	$s = 2$ (q_2)	...	$s = S$ (q_S)
G	x_1^G	x_2^G	...	x_S^G
B	x_1^B	x_2^B	...	x_S^B

Table 4.2 The state-space representation of risk.

The agent chooses a lottery θ. The payoff generated by the chosen lottery is exclusively enjoyed by a principal. The principal ("she", indexed by $p \in \{1, ..., P\}$) has perfect information. She is fully informed about the agent's choice as well as the realized state of the world. Note that the state-space description of uncertainty implies that the principal always observes the payoffs yielded by both lotteries. The principal then decides whether or not to award a fixed bonus, normalized to 1, to the agent or a passive player. That is, she chooses $bonus \in \{0, 1\}$, where $bonus = 1$ signifies that the principal awards the bonus to the agent, and $bonus = 0$ signifies that she awards the bonus to the passive player. There is no cost attached to this decision. Interactions are one-shot, meaning that there are no repeated play considerations.

4.2.2 Outcome biased perception of the agent's choice

The model is built to capture key patterns documented in the existing literature in a simple and parsimonious way. Principals are found to be more inclined to reward agents when 1) they chose the principal's preferred action (Charness and Levine, 2007; Gurdal et al., 2013; Rubin and Sheremeta, 2016; Davis et al., 2017; de Oliveira et al., 2017; Brownback and Kuhn, 2019; König-Kersting et al.,

2021); 2) conditional on the choice, the agent obtained a high outcome for the principal (Gurdal et al., 2013; de Oliveira et al., 2017; Brownback and Kuhn, 2019; König-Kersting et al., 2021) 3) conditional on the agent's choice and the obtained outcome, the non-chosen option yields a low outcome (Gurdal et al., 2013). These patterns are observed in one-shot interactions, and even when paying a bonus is costly to the principal. This suggests that motivations beyond material self-interest drive these patterns. Note that, since interactions are one-shot, rational self-interest does not provide any predictions about the principals' behavior in the considered setting. The broad idea of the model is that the principal has a desire to reward the agent for good choices, but the OB tricks the principal into evaluating the agent's choice as if he could have foreseen the realized state of the world.

Suppose the agent chooses lottery θ and state s materializes. In order to evaluate the agent's choice, the principal assesses whether he chose her preferred lottery. I assume that the principal's ex-ante preferences satisfy expected utility theory. She values money according to a strictly increasing utility function $u(\cdot)$ and has an ex-ante preference for lottery G whenever $\sum_{s=1}^{S} q_s u(x_s^G) \geq \sum_{s=1}^{S} q_s u(x_s^B)$. After having observed the realized outcome, she attempts to compute the expected utility of both lotteries but does so in an outcome-biased way. In modeling this, I follow an argument made prominently in the psychology literature. Baron and Hershey (1988) argue that OB might occur because it induces observers to focus too much on the realized outcomes when evaluating decisions. To formalize this idea, I assume that, after having observed the realization of state s, the principal inflates the decision weight put on the realized state of the world and deflates the weights put on the non-realized states. The principal's outcome-biased expected utility of lottery θ in state s, $OBEU_s^\theta$, is given by:

$$OBEU_s^\theta = \underbrace{[q_s + \lambda_p(1-q_s)]}_{\text{Overweighting materialized state}} u(x_s^\theta) + \sum_{j \neq s} \underbrace{(1-\lambda_p)q_j}_{\text{Underweighting other states}} u(x_j^\theta), \tag{4.1}$$

where $\lambda_p \in [0,1]$ captures the principal's degree of OB. For $\lambda_p = 0$, the weight put on the realized state is equal to the objective probability q_s. In this case, the expression in equation 4.1 simply collapses to the unbiased expected utility of lottery θ. For $\lambda_p = 1$, full weight is put on the realized state and the outcome-biased expected utility is given by the realized utility. As the realized state is overweighted, the non-realized states are proportionally underweighted by $(1-\lambda_p)$. This ensures that all decision weights sum up to 1. Given that states do not have an inherent meaning in the considered setting, one might argue that the uniform underweighting of non-realized states is the most natural assumption. With the suggested formulation, the distortion of decision weights is decreasing in q_s.

The above can be conveniently rewritten as

$$OBEU_s^\theta = \lambda_p \underbrace{u(x_s^\theta)}_{\text{Realized outcome}} + (1-\lambda_p) \underbrace{\sum_{s=1}^{S} q_s u(x_s^\theta)}_{\text{Ex-ante expected utility}} \tag{4.2}$$

Thus, the outcome-biased expected utility is the sum of the ex-post utility derived from the realized outcome and the ex-ante expected utility of lottery θ, weighted by the principal's degree of OB λ_p.[6]

To assess the agent's lottery choice, the principal compares the outcome-biased expected utility of the chosen lottery θ with that of the non-chosen lottery, denoted by $-\theta$. If the principal is driven by a desire to reward the agent for good choices and not for bad ones, her bonus decisions adhere to the following rule.

$$bonus = 1 \iff \lambda_p \underbrace{[u(x_s^\theta) - u(x_s^{-\theta})]}_{\text{Ex-post comparison of outcomes}} + (1-\lambda_p) \underbrace{[EU^\theta - EU^{-\theta}]}_{\text{Ex-ante comparison of EU}} > 0, \tag{4.3}$$

where $EU^\theta = \sum_{s=1}^{S} q_s u(x_s^\theta)$. That is, the principal's bonus decision is driven by two components. On the one hand, she compares the realized utilities of the two lotteries. On the other hand, she compares the difference in ex-ante expected utilities. These two components are weighted by the principal's degree of OB λ_p.[7]

In the appendix (section B.2.1), I show that expression 4.3 can be obtained by injecting the outcome-biased perception of the agent's lottery choice into a model of reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004). The principal's reciprocity motive induces a desire to award the bonus whenever she perceives that the agents is kind to her. The outcome-biased principal perceives the agent as kind whenever condition 4.3 holds, and not otherwise.

A brief discussion of the modeling assumptions is in order at this point. First, expression 4.3 captures the key patterns of bonus decisions reported in the literature. That is, ceteris paribus, the probability that principals reward agents is weakly increasing in the ex-ante quality of choice and the realized outcome and weakly decreasing in the forgone payoff. Second, modeling OB as impacting bonus decisions through perceived choice quality is consistent with the OB effects on a diverse set of outcomes. These include choice quality (Baron and Hershey, 1988; König-Kersting et al., 2021), perception of their "type"/skill (Baron and Hershey, 1988; Brownback and Kuhn, 2019), and perceived intentions (Sezer et al., 2016).

[6] This formulation has similarities to models of hindsight bias (Biais and Weber, 2009) and Madarász (2012)'s model of information projection, in which decision makers' beliefs are a convex combination of the true value of an underlying variable and an observation made by the decision maker.

[7] To avoid technicalities, I do not consider the case in which equation 4.3 is satisfied with equality. The results stated in Proposition 3 obtain also if one replaces the strict with a weak inequality or if is awarded randomly in this case.

4.2.3 Perverse incentives

I now discuss how the principal's outcome-biased reward decisions shape the agent's implicit incentives. Note that expression 4.3 implies that, in a given state, the principal will award the bonus either for a choice of lottery θ or $-\theta$. It is thus sufficient to consider bonus decisions for lottery G. For notational ease, denote $\Delta_s = u(x_s^G) - u(x_s^B)$ and $\Delta(EU) = EU^G - EU^B$.

If the agent chooses lottery G, a principal with a degree of OB λ_p awards the bonus with probability $Pr(\lambda_p) = \sum_{s=1}^{S} q_s \mathbb{1}\{\lambda_p \Delta_s + (1-\lambda_p)\Delta(EU) \geq 0\}$, and with probability $1 - Pr(\lambda_p)$ if the agent chooses lottery B. The incentives to choose lottery G imposed by the principal's bonus decisions are given by $I(\lambda_p) = 2Pr(\lambda_p) - 1$. The principal's bonus decisions thus impose incentives for the agent to choose lottery G whenever $Pr(\lambda_p) \geq 0.5$.

Proposition 3 collects the main results on how OB impacts the implicit contract between agent and principal. It establishes conditions under which OB can induce the principal to put in place incentives that effectively incentivize the agent to choose the principal's least preferred lottery.

Proposition 3. *Suppose $EU^G > EU^B$.*

a) $I(\lambda_p)$ is weakly decreasing in λ_p;

b) Whenever $\sum_{s=1}^{S} q_s \mathbb{1}\{\Delta_s < 0\} > 0.5$, $I(\lambda_p) < 0$ for all $\lambda_p \in (\underline{\lambda_p}, 1]$;

c) For any lottery G with at least 3 distinct outcomes and $\max_s\{q_s\} < 0.5$, there exists a lottery B, such that G is first-order-stochastic dominant to B, but $I(\lambda_p) < 0$ for all $\lambda_p \in (\underline{\lambda_p}, 1]$.

The proof is relegated to appendix AB.2.2.

Statement a) is a direct implication of the modeling assumptions. Intuitively, whenever the principal's preferred lottery G yields a higher payoff than lottery B, that is, when $\Delta_s > 0$, the principal awards the bonus for lottery G, but not B, for all $\lambda_p \in [0, 1]$. Whenever lottery G yields a lower payoff than lottery B, that is $\Delta_s < 0$, the principal might award the bonus if her degree of OB λ_p is above a threshold level, that is if $\lambda_p > \frac{\Delta(EU)}{\Delta(EU)-\Delta_s}$. Thus, the higher the principal's degree of OB, the more likely she is to reward the agent for choosing her least preferred lottery.

Statement b) shows that OB can induce a sufficiently outcome-biased principal to set perverse incentives, that is, incentives that contradict her ex-ante preferences. Note that for $\lambda_p = 1$, the principal awards the bonus solely based on the ex-post outcome comparison, that is $Pr(\lambda_p) = \sum_{s=1}^{S} q_s \mathbb{1}\{\Delta_s < 0\}$. Therefore, whenever lottery G is more likely to yield a higher payoff state-by-state, the principal is more likely to award the bonus if the agent chooses lottery B. More generally, there exists a threshold value of OB, $\underline{\lambda_p}$, such that a sufficiently outcome-biased principal is more likely to award the bonus for a choice of the least preferred lottery B if it is more likely to yield a higher outcome in the ex-post comparison. This threshold value can be defined as follows. Order the S states of the world such that $\Delta_1 \leq \Delta_2 \leq ... \leq \Delta_S$. Define state k such that $\Delta_k \in \{\max_s : \sum_{s=1}^{k} q_s > 0.5\}$. Then, $\underline{\lambda_p} = \frac{\Delta(EU)}{\Delta(EU)-\Delta_k}$.

Statement c) establishes conditions under which perverse incentives can occur even if the preferred lottery G is first-order stochastic dominant to lottery B. Given statement b), it suffices

to show that for every lottery G satisfying the stated conditions, one can find a lottery B that is 1) first-order-stochastically dominated by G, and 2) is more likely to yield a higher payoff in the ex-post comparison. In the appendix (section AB.2.2) I show that for a lottery G with K distinct outcomes, $max \sum_{s=1}^{S} q_s \mathbb{1}\{\Delta_s < 0\} = 1 - max_k\{q_k\}$, which yields the result immediately.

Statement c) also highlights the fundamental reason why OB can induce an agency problem. The incentives put in place by the principal depend not just on the marginal distribution of each lottery but also on how payoffs are correlated across states. This is exemplified in the tasks displayed in Table 4.1 in the introduction. The marginal distribution is the same under both correlation structures, and the principals always prefer the dominant lottery but the incentives set by a sufficiently outcome-biased principal change with the correlation structure. As such, the model of OB shares some similarities with correlation-sensitive models of risk preferences (Loomes and Sugden, 1982; Bell, 1982; Bordalo et al., 2012; Lanzani, 2022). In particular, the result in Proposition 3c is similar to Proposition 3 in Quiggin (1990), which provides conditions under which a dominated lottery might be preferred by a regret-averse individual. In appendix AB.2.3, I provide a formal argument to show that the sensitivity of incentives to the correlation structure is the reason why the principals' preferences and the provided incentives diverge.

However, there are two key differences between the proposed model of OB decision-theoretic work on correlation sensitivity. First, regret (Loomes and Sugden, 1982; Bell, 1982; Loomes and Sugden, 1987) and salience theory (Bordalo et al., 2012), the two prominent models that induce correlation-sensitivity of preferences, do not predict violations of first-order-stochastic-dominance for any lottery task that satisfies the assumptions of Proposition 3c.[8] Second, in regret and salience theory, it is ex-ante preferences that are correlation-sensitive. That is, correlation-sensitivity in choices (for oneself) may satisfy ex-ante preferences. OB, on the other hand, can induce the principal to set incentives that are correlation-sensitive, even if her ex-ante preferences are not correlation-sensitive.

I take the view that the principal's ex-ante preferences reflect her true preferences, whereas the incentives induced by OB are the result of cognitive bias, essentially a mistake made by the principal. In section 4.6 I present some evidence suggesting that this is indeed the case. This view implies that the basis for the evaluation of the principal's welfare are her ex-ante preferences and not her ex-post bonus decisions. The view that mistakes can produce behavior that should not be used as a standard for welfare is consistent with arguments made in Bernheim and Rangel (2009) and Bernheim (2021).

4.2.4 The agent's choices

I close the description of the model with a discussion of the agent's side. The agents' incentives to choose lottery G are given by $I(G) = \int (2Pr(\lambda_p) - 1) dH(\lambda_p)$, where $H(\lambda_p)$ is the cumulative

[8] More generally, the kind of correlation-sensitivity implied by regret and salience theory produces choice patterns opposite to the ones implied by the model OB. Roughly speaking, both regret and salience theory can be thought of as over-weighting states with a large payoff difference. In the extreme, a decision maker with such preferences might choose between lotteries by considering only the state with the largest payoff difference. OB, on the other hand, implies that incentives favor the lottery, which yields a higher payoff in most states.

distribution function of outcome bias in the relevant population. Agents have no information about the degree of OB of their principal but they have to form beliefs. As a benchmark, it is convenient to assume that the agent's beliefs are correct, that is $E[I(G)] = I(G)$, and that the agent chooses a lottery θ in order to maximize his monetary payoff. That is, the agent chooses lottery G whenever $E[I(G)] \geq 0$ and B otherwise. Proposition 1b) and especially c) thus pin down conditions under which the principal's OB, when anticipated correctly, can induce agents to take actions that are suboptimal from the point of view of the principal. Allowing for agents to be intrinsically motivated to make a good choices, for instance because of altruism does not change the main predictions presented below qualitatively. The possibility of other regarding preferences will be discussed in section 4.5.

Note that the correct expectations assumption implies that the agent knows the joint distribution over the principals' degree of OB, as well as the parameters of her utility function. Experimentally, this issue can be largely circumvented (see section 4.3.2 for details). Regarding the anticipation of the principal's OB, there is some evidence suggesting that individuals might even be better at predicting the biases of others than their own biases (Pronin et al., 2002; Pronin, 2007). However, it should be stressed that it is part of the goal of the experiment to test whether or not agents are able to form correct expectations in this setting.

4.3 Experimental Design and Hypotheses

4.3.1 Design

The experimental design follows the setting outlined above closely. The agent makes a series of lottery choices. The payoff of the lottery goes to his principal. The principal observes the agent's choice and, depending on the treatment, the realized state of the world. Then she decides whether or not to award a fixed bonus of 10 Euros to the agent. Agents and principals are randomly and anonymously matched. Interactions are thus one-shot. Similar to a design feature of Gurdal et al. (2013), the principal has to give away the bonus in any case and decides whether to give it to a randomly chosen participant in the role of an agent if the principal decides not to reward the matched agent. There is no commutation and no ex-ante contract. The setting thus captures important features of environments with no or incomplete contracts, in which agents have to form beliefs about their incentives to choose between different actions.

The structure of the choice tasks is inspired by proposition 3c. Consider Table 4.3. The agent chooses a lottery θ_k^{corr}. Choices are always between a dominant lottery G ("Good") and a dominated lottery B ("Bad"), that is $\theta \in \{G, B\}$. Each lottery yields a high payoff h, a medium m, and a low payoff l with a probability of 1/3 each. In addition, the dominant lotteries yield a payoff premium $\Delta > 0$ in each state. I employ four lottery pairs with different values of h, m, l, and Δ indexed by $k \in \{1, 2, 3, 4\}$. In a within-subject treatment, participants encounter each lottery pair in two correlation structures denoted by $corr \in \{1s, 2s\}$. Payoffs are chosen such that, under correlation structure $1s$, the dominant lottery G_k^{1s} does better than the dominated alternative in

only 1 out 3 states ("1s"), whereas the dominated lottery yields a higher outcome in 2/3 states. Under correlation 2s, it is the dominant lottery G_k^{2s} that yields a higher payoff in 2 out 3 states ("2s"). From proposition 3b, it is immediate that a sufficiently outcome-biased principal is more likely to award the bonus after an agent chose the dominated lottery under correlation structure 1s but never under correlation structure 2s.

The parameterization of the four lottery pairs is shown in Table 4.4. Payoffs are varied systematically such that the minimal λ_p required for a principal to incentivize agents to choose the dominated lottery B is increasing from choice pair 1 to 4.[9] This design allows both gauging the strength of OB non-parametrically as well as identifying principals' degree of OB using structural estimation. An additional choice task was added that features only correlation 1s. This choice task was added solely to aid with the structural estimation.

1s: $G > B$ in 1/3 of states			
	1(1/3)	2(1/3)	3(1/3)
G_k^{1s}	$h+\Delta$	$m+\Delta$	$l+\Delta$
B_k^{1s}	l	h	m

2s: $G > B$ in 2/3 of states			
	1(1/3)	2(1/3)	3(1/3)
G_k^{2s}	$h+\Delta$	$m+\Delta$	$l+\Delta$
B_k^{2s}	m	l	h

Table notes: Parameters satisfy $h > m > l$, $\Delta > 0$, $h > m + \Delta$, and $m > l + \Delta$.

Table 4.3 The structure of the choice tasks employed in the principal-agent interactions of the experiment.

Lottery pair (k)	h	m	l	Δ
1	1953	1031	109	45
2	1953	1031	109	110
3	1403	688	103	359
4	1403	688	103	523
5 ("1s" only)	1480	750	50	699

Table 4.4 Parameters of the choice task used in the experiment.

In addition to the within-subject variation described above, there are two between-subject treatments. In the *reward-after* treatment, principals can condition their bonus payments on choices and outcomes. In this treatment, bonus decisions can be influenced by OB. In the *reward-before* treatment, principals make their bonus decisions before being informed about the outcome of the lotteries. Hence, OB cannot influence bonus decisions in this treatment. In the model discussed above, this can be seen as enforcing $\lambda_p = 0$.

4.3.2 Risk preferences are controlled for

A number of design choices deserve further discussion. First, while having agents choose between dominant and dominated lotteries allows for a tight link of the design with the Proposition 3, this feature also serves additional purposes. Most importantly, using first-order-stochastic-dominance implies that there is one lottery that is optimal and should be preferred by principals regardless of their risk preferences. When choosing on behalf of the principals, agents need not anticipate the principals' risk preferences.[10] It should be clear to agents which lottery the principals prefer.

[9] Assuming linear utility, a principal will incentivize agents to choose lottery B if $\lambda_p > 0.049$ for lottery pair 1, $\lambda_p > 0.120$ for lottery pair 2, $\lambda_p > 0.500$ for lottery pair 3, $\lambda_p > 0.900$ for lottery pair 4.

[10] Imposing that the principal's valuation function is described by EUT, the principals' bonus decisions are relatively unaffected by the curvature of their utility function unless risk preferences are extreme.

Finally, using dominant lotteries also facilitates drawing conclusions about the welfare implications of OB.

Second, the binary nature of bonus decisions allows for a clear interpretation of the agents' incentives. Binary bonus decisions imply that the agents' incentives to choose the dominant lottery boil down to the difference in receiving the bonus when choosing the dominant rather than the dominated lottery. Any self-interested agent should choose the lottery for which they expect a higher likelihood of receiving the bonus, regardless of their risk preferences.[11] Further, forcing principals to give the bonus to a random agent whenever they do not give it to the matched agents controls for wealth effects, efficiency concerns, and risk-sharing motives on the principal's side (Gurdal et al., 2013). The bonus decision thus arguably captures a desire to reward or punish the agent.

4.3.3 Identification of outcome bias

Most importantly, the chosen experimental setting allows for a clean identification of the effect of OB on bonus decisions and incentives. Changing the correlation structure allows to change the counterfactual comparison while keeping the marginal distribution of the lotteries constant. This implies that the outcome obtained by the principal is the same under both correlation structures. The identification of OB relies solely on variation in the forgone outcome.

The design thus allows distinguishing the effects of OB from other channels that could render bonus decision outcome dependent, such as outcome-based reciprocity, a risk-sharing motive, or distributional preferences. To see this, consider state 2 of both correlation structures displayed in Table 4.3. If an agent chooses the dominant lottery, the principal receives a payoff of $m + \Delta$ under both correlation structures. Models of distributional preferences or outcome-based reciprocity thus predict equivalent bonus decisions. However, under correlation structure $1s$, the dominated lottery yields a payoff $h > m+\Delta$, whereas it yields the low payoff $l < m+\Delta$ under correlation structure $2s$. Therefore, the model of OB predicts that principals might not award the bonus under correlation structure $1s$, but might do so under correlation structure $2s$.

The reward-before treatment addresses the concern that bonus decisions might change from one correlation structure to the other for reasons other than OB. In the reward-before treatment, principals cannot condition their bonus decisions on the realized state. Therefore, changing the correlation structure can only influence bonus decisions through channels other than OB. Comparing how changing the correlation structure impacts bonus decision in the reward-before and the reward-after treatment allows isolating effects that are due to ex-post counterfactual comparison. Moreover, to gauge whether principals' ex-ante preferences are impacted by the change in the correlation structure, principals are asked to choose between the lotteries used in the principal-agent interaction for themselves.

[11] Note that allowing principals to make more continuous bonus decisions would likely imply bonus payments that differ in size across states. The agent's lottery choice, again in terms of monetary rewards, would thus be a choice between different lotteries. Agents with different risk preferences might prefer different lotteries, which would make the interpretation of their incentives more difficult.

4.3.4 Main hypotheses

In the following, I collect the main experimental hypotheses. All of these hypotheses were preregistered with the AER social science registry under ID AEARCTR-0011213.[12] They derive from the model of OB discussed above. Hypothesis 4 concerns the principals' bonus decision.

Hypothesis 4. ***Outcome bias in bonus decisions:*** *Principals are more likely to award the bonus if*

a) Agents chose their preferred lottery, as measured by their own choices at the end of the experiment.

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

Hypothesis 5 concerns how the principal's bonus decisions shape the implicit incentives agents have to choose the dominant lottery. Denote by $I_p(G_k^{corr}, treatment)$ the incentives to choose the dominant lottery under correlation $corr \in \{1s, 2s\}$, in $treatment \in \{before, after\}$ as set by principal p's bonus decisions. These incentives are given by $I_p(G_k^{corr}, treatment) = Pr_p(G_k^{corr}, treatment) - Pr_p(B_k^{corr}, treatment)$, where $Pr_p(\theta_k^{corr}, treatment)$ is the probability of awarding the bonus if an agent chose lottery θ.[13]

Hypothesis 5. ***Outcome bias and incentives:***

a) $I_p(G_k^{corr}, before) > 0$. *Moreover, no significant difference arises between* $I_p(G_k^{1s}, before)$ *and* $I_p(G_k^{2s}, before)$, *for all lottery pairs and both correlation structures.*

b) $I_p(G_k^{corr}, before) > I_p(G_k^{corr}, after)$, *for all lottery pairs and both correlation structures.*

c) $I_p(G_k^{1s}, before) - I_p(G_k^{2s}, before) > I_p(G_k^{1s}, after) - I_p(G_k^{2s}, after)$, *for all lottery pairs.*

Hypothesis 5 a) states that, in the reward-before condition, incentives to choose the dominant lottery are expected to be positive and are not expected to differ significantly between correlations 1s and 2s. This follows from imposing $\lambda_p = 0$ in the model. Hypothesis 5 b) states that OB is expected to decrease the incentives to choose the dominant lottery under both correlation structures. This follows directly from proposition 3a). Finally, hypothesis 5 c) derives from proposition 3c). It states that changing the correlation structure from $1s$ to $2s$ reduces the incentives to choose the dominant lottery more in the reward-after than the reward-before treatment. By comparing correlation effects between treatments, the channel of ex-post counterfactual comparison can be isolated. The difference-in-difference comparison is equivalent to testing whether moving from the reward-before to the reward-after treatment reduces incentives more under correlation $2s$ than $1s$.

Hypothesis 6. ***The agent's beliefs: The agents form correct beliefs regarding the principals' bonus decisions.***

[12] The pre-analysis plan can be found in the appendix AB.6 and can be accessed here: https://doi.org/10.1257/rct.11213-1.1.

[13] In the reward-after treatment, bonus probabilities are averaged over all states, that is $Pr_p(\theta_k^{corr}, after) = \sum_{s\in\{1,2,3\}} Pr_p(\theta_k^{corr}, after, s)$.

Hypothesis 6 concerns the agents' beliefs. The working hypothesis is that agents hold correct beliefs about the principals' bonus decisions. As detailed in the preregistration, this analysis is somewhat exploratory in nature, and no specific test was preregistered. Beliefs will naturally deviate from the principals' bonus decisions sometimes, and any criteria for how much deviation should still be regarded as implying correct beliefs are bound to be subjective. In the analysis, I will employ a multi-pronged approach to address this question.

Hypothesis 7 collects the main hypotheses on the agents' choices. Denote $F(G_k^{corr}, treatment)$ the frequency with which agents choose the dominant lottery under correlation $corr \in \{1s, 2s\}$, for lottery pair $k \in \{1, 2, 3, 4\}$, and in $treatment \in \{before, after\}$.

Hypothesis 7. ***The agent's choices:***

a) In the reward-before treatment, agents choose the dominant lottery at a high frequency under both correlation structures. Moreover, no significant difference arises between $F(G_k^{1s}, before)$ *and* $F(G_k^{2s}, before)$.

b) $F(G_k^{corr}, before) > F(G_k^{corr}, after)$*, for all lottery pairs and both correlation structures.*

c) $F(G_k^{1s}, before) - F(G_k^{2s}, before) > F(G_k^{1s}, after) - F(G_k^{2s}, after)$*, for all lottery pairs.*

Hypothesis 7 mirrors Hypothesis 5. If Hypothesis 5 and 6 hold, agents are expected to choose the dominant lottery at a high frequency in the reward-before treatment under both correlation structures. In the reward-after treatment, the dominant lottery is expected to be chosen less frequently under both correlation structures. Finally, changing the correlation structure from 1s to 2s is expected to reduce the frequency of dominant lottery choices more in the reward-after than in the reward-before treatment.

The above hypotheses are tested for all four lottery pairs separately. Hypotheses 1b), 2, and 4 are most likely to be supported by the data for lottery pair 1, less likely for lottery pair 2, etc., as higher levels of OB are required for them to hold true.

4.4 Procedures

The experiment was conducted at the Lab of the Toulouse School of Economics with a total of 146 participants in the reward-after and 136 in the reward-before condition, and an equal number of principals and agents in each treatment. Participants were randomly assigned to the role of principal or agent. Each session was assigned either to the reward-before or reward-after treatment.

A total of 17 sessions were conducted, with a minimum of 6 and a maximum of 26 participants per session. The brunt of the sessions were carried out between the 13th and the 21st of April 2023. Two catch-up sessions were conducted on the 23rd of May. Sessions lasted between 45 and 75 minutes. Participants were paid a show-up fee of 5 Euros. Additional earnings ranged from 0 to 21.79 Euros, with final earnings averaging at 15.91 Euros. Participants were mostly French (77%), and undergraduate students (75%). 51.4% of the participants majored in Economics and

were 53.5% female. The experiment was programmed in oTree (Chen et al., 2016). Participants were recruited using ORSEE (Greiner, 2015).

The experiment consists of two parts. Part I concerns the principal-agent interaction. Part II includes a survey and, for principals only, a number of lottery choices they make for themselves. For sample screenshots, see section B.7.1. The complete instructions are provided here.

4.4.1 Part I: Principal-agent interaction

Principals: Principals make their bonus decisions in a variant of the strategy method (Selten, 1967). In the reward-before treatment, they decide whether or not to award the bonus to the agent conditional on the agent's lottery choice. In the reward-after condition, principals make their bonus decision for each possible choice outcome combination. Principals thus make a total of $2 * 9 * 3 = 54$ bonus decisions in the reward-after condition and $9 * 2 = 18$ bonus decisions in the reward-before condition. The strategy method was chosen because it allows obtaining a full picture of the principals' bonus decisions, which is crucial to understanding how OB impacts incentives.

Lotteries are described as generating payoffs that depend on the turn of a virtual wheel of fortune with sixty fields. The use of a wheel of fortune allows for a natural implementation of the correlation structures and also highlights the role of luck in determining the outcome of the agent's choice. In both treatments, principals see a table describing the lottery choice. Lotteries are neutrally labeled "Option A" or "Option B". In the reward-after treatment, principals make bonus decisions one state at a time, with the realized state highlighted in yellow. This design for the reward-after treatment was selected to mimic a situation in which principals are confronted with a particular outcome while retaining the benefits of using the strategy method.[14]

Agents: Agents make 2 choices for each choice task. For each choice task, they make a first choice prior to any belief elicitation. Thereafter, agents are informed about the belief elicitation. Agents in the reward-after condition were asked to state the expected probability with which they will receive the bonus conditional on choosing a lottery and a realized state. In the reward-before condition, agents were asked to state a bonus conditional on the lottery choice only. Agents thus stated a total of 54 beliefs in the reward-after and 18 beliefs in the reward-before condition. Directly after stating their beliefs for a given choice task, agents were asked to make a second choice for the respective task and then moved on to the next choice task.[15] Note that this design allows obtaining a first set of choices that are not influenced by the belief elicitation. It seems possible that the belief elicitation might induce agents to think through their choice more systematically, which is why the belief elicitation is followed by a second choice.

The belief elicitation is incentivized using the binarized scoring rule (Hossain and Okui, 2013).

[14] Principals encounter the reward decisions in random order. I further randomize the order in which states appear on the screen and which option is displayed at the top and the bottom row of the table describing the choice task.

[15] The order in which agents encounter the different tasks, as well as the display (order of states, position of lottery G and B), are randomized between subjects but are fixed for a given agent.

Instructions are simplified, following Danz et al. (2022).[16] Subjects were asked to imagine that they will be randomly matched with a principal in their session 100 times and to state the number of times expect to receive a bonus from the matched principal.[17] Agents entered their beliefs by moving a slider in the range of 0-100.[18]

Instructions for principals and agents differed only minimally. Principals and agents were neutrally referred to as orange and blue players, following Brownback and Kuhn (2019). In an effort to introduce common knowledge, a summary of the experimental instructions was read out aloud to participants at the beginning of the experiment. After this, participants proceeded to detailed instructions, which they read in silence. After completing the instructions, participants had to answer a set of comprehension questions. If a question was not answered correctly, feedback was given. Participants were allowed to proceed only once they had answered all comprehension questions correctly.

4.4.2 Part II: risk preference measurement and survey

In part II, principals make one choice for each of the nine choice tasks used in part I, now for themselves. In addition, they complete three multiple-choice lists with 11 lottery choices each. For a detailed discussion of these tasks, see appendix B.4.5.

Part II further included a number of survey items. These included an extended, seven-item cognitive reflection test (CRT, Frederick (2005); Toplak et al. (2014)),[19] and self-reported willingness to take risk (Dohmen et al., 2011). Moreover, subjects in the role of principal are asked to which extent 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 which extent their lottery choices were driven by 1) a desire to make good choices, 2) maximization of the probability of obtaining 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 chances of receiving the bonus.

Finally, the survey included a standard demographics questionnaire comprised of age, gender, field of study, nationality, level of education, and household income.

[16] Danz et al. (2022) suggests that providing detailed information about the binarized scoring rule might result in center-biased belief reports. I, therefore, follow their recommendation and inform subjects that the payment rule is designed in a way such that the most accurate guess is most likely to win a prize of 15 Euros, without providing specific details on the payoff mechanism. By clicking on a link, subjects could access detailed information about the payoff mechanism.

[17] This frequency rather than a probability framing was implemented because some research suggests that the former is often better understood (Gigerenzer and Hoffrage, 1995).

[18] To avoid potential biases due to the initial starting position of the slider, the initial position was masked in a design adapted from (Grossmann, 2023).

[19] The CRT consists of questions that have an intuitive, though wrong answer. The test is commonly used to gauge to which extent participants engage in cognitive reasoning rather than responding intuitively and is a good predictor of other cognitive measures (Frederick, 2005; Toplak et al., 2014). As the original three-question version of the test is quite popular, there is a worry that subjects might have previously encountered the test, which could reduce its validity (Haigh, 2016), but see Bialek and Pennycook (2018) for evidence to the contrary. To alleviate these concerns, the survey included the original three questions as well as four additional items proposed by Toplak et al. (2014). Subjects are further asked to indicate whether they have seen one of the questions before, separately for the original three items and the four-item extension.

4.4.3 Payments

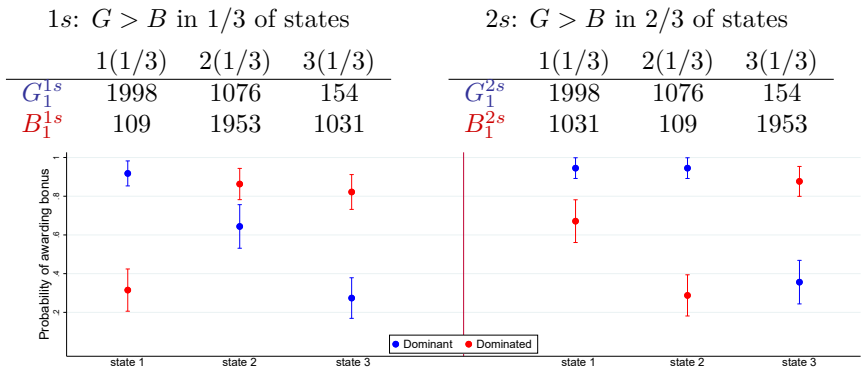
Participants were paid according to a random incentive mechanism to avoid potential issues arising from hedging beliefs and choices. With 80% chance, their additional payment was determined based on the principal-agent interaction. In this case, each agent was randomly and anonymously paired with a principal. One of the agent's choices was selected at random. The first and the second set of choices were equally likely to be chosen. The computer then randomly selected a state of the world, and the bonus decision specified by the principal was implemented. With 20% chance, agents were paid according to one randomly chosen belief, and principals were paid according to one randomly chosen lottery choice they made in part II. As behavior in the principal-agent part is somewhat more important for the purpose of this study, more weight was given to this part of the experiment.

4.5 Results

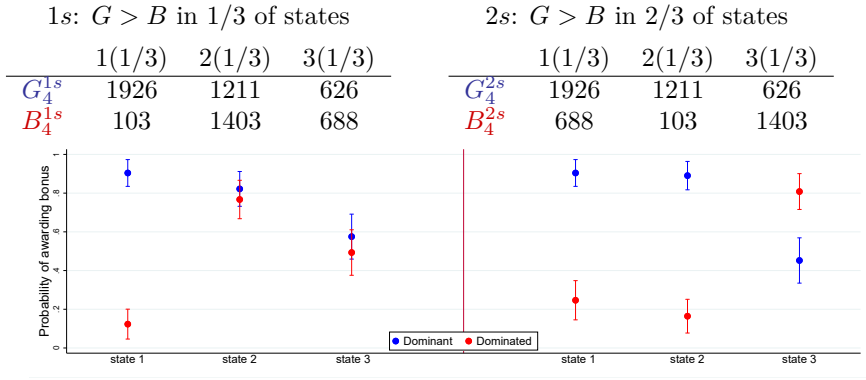
4.5.1 The principals' bonus decisions - Descriptives

Before delving into the principals' bonus decisions, a short discussion of their preferences, as revealed by the choices they made for themselves at the end of the experiment, is in order. Averaged over all lottery pairs and both treatments, principals chose the dominant lottery at a frequency of 94%. There is no evidence for treatment differences in lottery choices or that preferences are influenced by the change in the correlation structure. See section B.3.1.1 in the appendix for more details.

The bonus decisions of principals in the reward-before treatment reflect their preferences rather well. Averaging over all four lottery pairs, principals award the bonus for a choice of the dominant lottery with 85.7% probability. They reward a choice of the dominated lottery with a significantly lower frequency of 33.8% ($p < 0.001$ for all choice tasks, two-sided McNemar's exact test). As we will see in more detail below, bonus decisions are not significantly influenced by the change in the correlation structure.



(a) Lottery pair 1



(b) Lottery pair 4

Table notes: Bonus frequencies are averaged over the 73 principals in the reward-after treatment, with 95% confidence intervals. The states are indicated on the x-axis and are below the corresponding state in the table displaying the choice task.

Figure 4.1 Frequencies of awarding the bonus for a given choice-state combination.

Let us now turn to the reward-after treatment. Figure 4.1 depicts the probability with which principals in the reward-after condition award the bonus conditional on an agent's choice and the realized state for lottery pairs 1 and 4.[20] Absent the OB, one might expect principals to behave similarly to their counterparts in the reward-before treatment and reward choices of the dominant lottery, regardless of the realized state. Figure 4.1 shows that this is not the case. Principals have a strong tendency to condition their bonus decisions on the realized state. Consider lottery pair 1, for which the model predicts the strongest OB effects. The probability of awarding the bonus for a choice of the dominant lottery G_1^{1s} varies from 90% in state 1 to 24% in state 3. Likewise, the probability of awarding the bonus for a choice of the dominated lottery B_1^{1s} ranges from 83% in state 2 to 17% in state 1. Moreover, counterfactual comparison between the obtained and the forgone option seems to play an important role. Observe that, in each state, principals are significantly more likely to award the bonus for whichever lottery yielded a higher payoff ($p < 0.003$ for all comparisons, two-sided McNemar's exact test).

Comparing bonus decisions across the two lottery pairs reveals interesting patterns. Recall that the OB model implies that outcome effects are strongest for lottery 1 and weakest for lottery pair 4. The bonus decisions of principals are consistent with this prediction. For lottery pair 1, principals are significantly more likely to award the bonus for a choice of the dominated lottery in

[20] Throughout the paper, I illustrate the results mainly on these two lottery pairs because OB effects should be the strongest for pair 1 and the weakest for pair 4. I supply figures for lottery pairs 2 and 3 in the appendix, section B.3.

all three states in which it yields a higher payoff. For lottery pair 4, this only holds true only in one state.[21] Second, observe that even for lottery pair 4, for which the premium of the dominant lottery amounts to 71% of the expected value of the dominated lottery, bonus decisions are still strongly outcome-dependent. This suggests that the experimental subjects display rather strong OB.

4.5.2 OB in Bonus decisions - Preregistered hypotheses

I now turn to the testing hypothesis 4 on whether principals display OB. As pre-registered, I test the hypothesis by estimating the following random-effects logistic regression model for each of the four lottery pairs separately.

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

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 choice at the end of the experiment), $obtained\ payoff_{i,t}$ 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. As preregistered, one-sided p-values are reported whenever the preregistered alternative hypothesis has a clear direction ($<$ or $>$). Two-tailed p-values are reported otherwise.

[21] Bonus decisions for lottery pairs 2 and 3 align nicely with this pattern, with bonus decisions displaying an outcome dependence that seems intermediate to that in pairs 1 and 4. See figure 4.1.

	Reward-after			
	(1) Pair 1	(2) Pair 2	(3) Pair 3	(4) Pair 4
$Preferred$	0.053**	0.147***	0.189***	0.205***
	(0.026)	(0.027)	(0.030)	(0.035)
$obtained > forgone$	0.204***	0.155***	0.177***	0.161***
	(0.049)	(0.049)	(0.049)	(0.047)
$obtained\ payoff$ (in Euro)	0.019***	0.020***	0.028***	0.025***
	(0.003)	(0.003)	(0.004)	(0.004)
Observations	876	876	876	876
Individuals	73	73	73	73

	Reward-before			
	(1) Pair 1	(2) Pair 2	(3) Pair 3	(4) Pair 4
$Preferred$	0.377***	0.422***	0.593***	0.612***
	(0.052)	(0.052)	(0.055)	(0.064)
$obtained > forgone$	0.056	-0.053	0.005	-0.100
	(0.097)	(0.089)	(0.077)	(0.079)
$obtained\ payoff$ (in Euro)	-0.000	0.006	0.015**	0.010
	(0.006)	(0.006)	(0.007)	(0.007)
Observations	272	272	272	272
Individuals	68	68	68	68

Table notes: The top panel displays results for the reward-after and the bottom panel results for the reward-before treatment. Standard errors, clustered at the subject level, are in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 4.5 Average marginal effects from random-effects logistic regression.

Table 4.5 displays estimates of average marginal effects (AME).[22] The top panel displays results for the reward-after treatment, and the bottom panel results for the reward-before treatment. As preregistered, I test hypothesis 4a that principals are more likely to award the bonus if an agent chose her preferred lottery by testing whether $\beta_1 > 0$ against the null that $\beta_1 = 0$. For all lottery pairs and both treatments, the null is rejected at ($p < 0.021$, Wald Chi-Square test) in favor of the stated alternative hypothesis. This lends strong support to hypothesis 4a.

The intensity of principals' tendency to reward agents for choosing their preferred lottery varies strongly across the different lottery tasks and treatments and displays interesting patterns beyond the preregistered hypotheses. First, within both treatments, the estimated coefficient increases monotonically from lottery pair 1 to lottery pair 4. While principals in the reward-after treatment are only 5.3 percentage points (ppts) more likely to award the bonus when the agent chose their preferred lottery for lottery pair 1, this number increases to 20.5 ppts for lottery pair 4 ($p = 0.001$[23]). In the reward-before treatment, the estimated effect increases from 37.7 ppts for lottery pair 1 to 61.2ppts for lottery pair 4 ($p = 0.013$). This pattern suggests that principals adapt their

[22] Table B.2 in the appendix reports coefficients of the regression model.

[23] The p-values for comparisons across lottery pairs are obtained by running the following random-effects logit regression model on bonus decisions for lottery pair 1 and 4, separately for the reward-before and reward-after treatment, and testing the null that $\beta_4 = 0$. $Bonus_{i,t} = \beta_0 + \beta_1 preferred_{i,t} + \beta_2\{obtained > forgone\}_{i,t} + \beta_3 obtained\ payoff_{i,t} + lp4_{i,t} * [\beta_4 preferred_{i,t} + \beta_5\{obtained > forgone\}_{i,t} + \beta_6 obtained\ payoff_{i,t}] + \beta_7 lp4_{i,t} + \epsilon_{i,t}$, where $lp4_{i,t}$ is a dummy indicating whether principal i made bonus decision t for lottery pair 4.

bonus decisions both to the extensive margin, i.e. whether an agent chose their preferred lottery, as well as the intensive margin, i.e. how much better the chosen lottery is. Second, for each lottery pair, bonus decisions react much stronger to the agent's choice in the reward-after than in the reward-before treatment ($p < 0.001$[24]). That is, the OB seems to crowd out the principals' tendency to reward the agents based on their choices. This finding confirms results of Brownback and Kuhn (2019). Note that both patterns can be explained by the proposed model of OB.

Hypothesis 4b states that principals are more likely to award the bonus when the obtained payoff is higher than the forgone payoff. As preregistered, I test this hypothesis by testing whether $\beta_2 > 0$ against the null that $\beta_2 = 0$ in the reward-after treatment. For all lottery pairs, the null is rejected at $p < 0.001$ (one-sided Wald Chi-Square test). Reassuringly, none of the estimated coefficients for the reward-before treatment is statistically significantly different from 0 ($p > 0.201$, two-sided Wald Chi-Square test). These results confirm hypothesis 4b. The point estimates of the AMEs suggest that principals are between 15.5ppts and 20.4 ppts more likely to award the bonus if the obtained outcome is higher than the forgone outcome. In the reward-after treatment, the effect of obtaining a higher outcome has a significantly stronger impact on the bonus probability than the agent's choice for lottery pair 1 ($p = 0.006$, chi-square test), and is not significantly different from the effect of the agent's choice for the other lottery pairs ($p > 0.299$, chi-square test). The result that principals condition bonus payments on the ex-post counterfactual comparison of outcomes confirms previous findings of Gurdal et al. (2013).

Finally, principals in the reward-after condition show a strong tendency to reward agents based on the realized payoff. All estimated AMEs are positive and statistically significant at $p < 0.003$. Estimates range from a 1.9ppts increase in the probability of awarding a bonus for an additional Euro obtained by the principal for lottery pair 1 to 2.9ppts for lottery pair 3.[25] Given that payoffs vary from 1.03 to 20.63, the estimated effect size is considerable. In the reward-before treatment, the estimated coefficients are not significantly different from 0 ($p > 0.17$), except for lottery pair 3, for which the estimate is positive and statistically significant at $p = 0.044$. As principals were unable to condition their bonus decisions on the outcome of the lottery, this is likely to be a false positive.

I summarize the results on the determinants of the principals' bonus decisions below.

Result 4. ***Outcome bias in bonus decisions***

a) In both the reward-after and the reward-before treatment, principals are significantly more likely to award the bonus if the agent chose their preferred lottery for all lottery pairs. The results confirm hypothesis 4a.

b) Principals in the reward-after treatment are between 15.5ppts for lottery pair 2 and 20.4ppts for lottery pair 3 more likely to award the bonus when the payoff obtained by the agent is higher than the forgone payoff. I confirm hypothesis 4b.

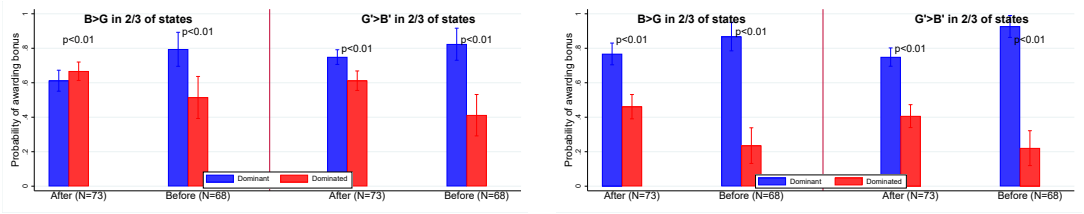
[24] P-values are obtained by running the following random-effects logit regression model, and testing whether $\beta_4 = 0$, for each lottery pair separately. $Bonus_{i,t} = \beta_0 + \beta_1 preferred_{i,t} + \beta_2\{obtained > forgone\}_{i,t} + \beta_3 obtained\ payoff_{i,t} + before_i * [\beta_4 preferred_{i,t} + \beta_5\{obtained > forgone\}_{i,t} + \beta_7 obtained\ payoff_{i,t}] + \beta_7 before_i + \epsilon_{i,t}$, where $before_i$ is a dummy indicating whether principal i is in the reward-before or reward-after treatment.

[25] None of the estimates differ from each other at $p < 0.05$.

4.5.3 Incentives to choose the dominant lottery

I now turn to how the principals' OB impacts the agents' incentives to choose the dominant lottery. Figure 4.2 displays the probability that principals award the bonus, conditional on the agents' lottery choice, for the reward-before and -after treatment for lottery pairs 1 and 4. Incentives follow a clear pattern.[26] In the reward-before treatment, agents have strong incentives to choose the dominant lottery under both correlation structures for both pairs. The incentives do not change significantly with the correlation structure (see below for more detail). In the reward-after treatment, the incentives to choose the dominant lottery are considerably reduced, relative to the reward-before treatment. Moreover, the incentives are lower when it is the dominated rather than the dominant lottery that yields a higher payoff in 2/3 states.[27] Consistent with the result that principals' tendency to reward agents based on their choices increases with the premium of the dominant lottery, incentives to choose the dominant lottery are larger for lottery pair 4 than 1.

Importantly, OB can indeed induce perverse incentives. For lottery pair 1, principals award the bonus with 61% probability if an agent chose the dominant lottery G_1^{s1}, but with 67% probability if an agent chose the dominated lottery B_1^{s1}. The difference of 6ppts is small but statistically significant at $p = 0.005$ (two-sided Wilcoxon signed-rank test[28]), which indicates that agents have small, albeit significant, incentives to choose the dominated lottery when it is more likely to yield a higher payoff than the dominant one. At the individual level, this implies that a significant fraction of the principals has revealed preferences for the dominant lottery but nevertheless provides incentives for agents to choose the dominated lottery.[29]



(a) Lottery pair 1 **(b)** Lottery pair 4

Table notes: With 95% confidence intervals. For the reward-after treatment, the frequencies are averaged over the three states.

Figure 4.2 Frequencies of awarding the bonus, conditional on the agent's lottery choice.

I now turn to the preregistered hypotheses. The first part of hypothesis 5 states that principals in the reward-before treatment provide strictly positive incentives to choose the dominant lottery and that incentives to choose the dominant lottery do not differ significantly across the correlation

[26] Incentives for lottery pairs 2 and 3 display similar patterns. See figure B.2.

[27] This is significant at $p < 0.05$ for lottery pairs 1-3 (two-sided Wilcoxon signed-rank test, see table in the appendix B.3 for details).

[28] Note that $Pr_p(\theta_k^{corr}, after) \in \{0, 1/3, 2/3, 1\}$, $Pr_p(\theta_k^{corr}, after) \in \{0, 1\}$, $I_p(G_k^{corr}, after) \in \{-1, -2/3, ..., 1\}$, and $I_p(G_k^{corr}, before) \in \{-1, 0, 1\}$. For this reason, I employ non-parametric tests in this section.

[29] 22 out of the 73 principals in the reward-after treatment are strictly more likely to award the bonus when the agent chose the dominated lottery B_1^{s1}, although, at the end of the experiment, they choose the dominant lottery G_1^{s1} for themselves. Only 6 of their 68 counterparts in the reward-before treatment display this behavior, which is a significantly lower fraction ($p = 0.002$, two-sample test of proportions). Further, only 3 of the principals in the reward-after and 4 principals in the reward-before treatment display a similar inconsistency when the dominant lottery yields a higher payoff in 2/3 of the states. This strongly suggests that it is indeed OB that induces some principals to put in place perverse incentives.

structures. I test this hypothesis for all lottery pairs separately, using Wilcoxon signed-rank tests. For all four lottery pairs, I reject the null hypothesis that $Pr_p(G_k^{corr}, after) = Pr_p(G_k^{corr}, before)$ in favor of the alternative hypothesis that $Pr_p(G_k^{corr}, after) > Pr_p(G_k^{corr}, before)$ at $p < 0.001$.[30] Moreover, I do not reject the null hypothesis that $I_p(G_k^{1s}, before)$ differs significantly from $I_p(G_k^{2s}, before)$ for any of the four lottery pairs ($p > 0.20$). These results confirm hypothesis 5a.

Hypothesis 5b states that incentives to choose the dominant lottery are lower in the reward-before than in the reward-after treatment for both correlation structures. I test the null hypotheses that $I_p(G_k^{corr}, before) = I(G_k^{corr}, after)$ against the alternative hypotheses that $I_p(G_k^{corr}, before) > I(G_k^{corr}, after)$, using Wilcoxon rank-sum tests. For all lottery pairs and both correlation structures, I reject the null hypothesis at $p < 0.002$. These results confirm hypothesis 5b.

Hypothesis 5c states that $I(G_k^{1s}, before) - I(G_k^{2s}, before) > I(G_k^{1s}, after) - I(G_k^{2s}, after)$. In words, changing the correlation structure from the one in which the dominated lottery yields a higher payoff in 2/3 of the states to the one in which it is the dominant one that yields a higher payoff more often is expected to increase the incentives to choose the dominant lottery more in the reward-after than in the reward-before treatment. I test this hypothesis against the null that $I(G_k^{1s}, before) - I(G_k^{2s}, before) = I(G_k^{1s}, after) - I(G_k^{2s}, after)$, for all lottery pairs separately using Wilcoxon rank-sum tests. The null hypothesis is rejected at $p = 0.063$ for lottery pair 1, $p = 0.085$ for lottery pair 2, $p = 0.002$ for lottery pair 3, and $p = 0.893$ for lottery pair 4. Overall, this lends some support to hypothesis 5c, although the evidence is not as strong as for the other parts of hypothesis 5.

I summarize the results on how OB shapes the incentives to choose the dominant lottery.

Result 5. ***Outcome bias and incentives:***

a) In the reward-before treatment, principals provide strong incentives to choose the dominant lottery, for all lottery pairs. The incentives to choose the dominant lottery do not change significantly with the correlation structure for any lottery pair.

b) For all lottery pairs, incentives to choose the dominant under both correlation structures are significantly lower in the reward-after than in the reward-before treatment.

c) In the reward-after treatment, changing the correlation structure from the one under which the dominated lottery yields a higher outcome in 2/3 states to the correlation structure in which it is dominant lottery that yields a higher payoff in 2/3 states increases the incentives to choose the dominant lottery more in the reward-after than in the reward-before treatment, This is statistically significant at $p = 0.063$ for lottery pair 1, $p = 0.085$ for lottery pair, and $p = 0.001$ for lottery pair 3. For lottery pair 4, there is no significant difference at any conventional significance level.

[30] I provide an overview of all the relevant test statistics for hypothesis 5 in table B.3.

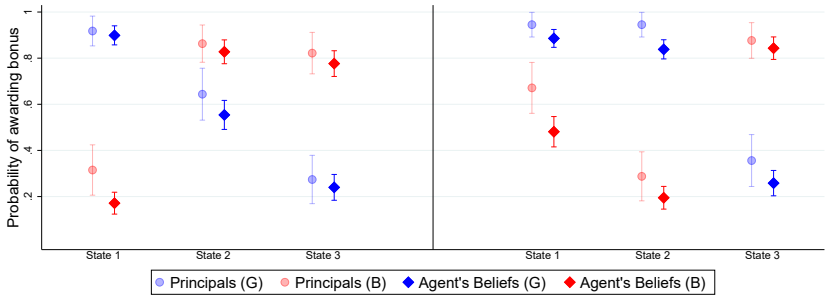
4.5.4 The agents' beliefs

I next turn to the agent's beliefs about he principals' bonus decisions. In the reward-before treatment agents are capable of anticipating the broad patterns in the principals reward-decisions. For each lottery pair, agents expect to be significantly more likely to receive the bonus when choosing the dominant rather than the dominated lottery. See appendix B.3.1.3 for details and discussion.

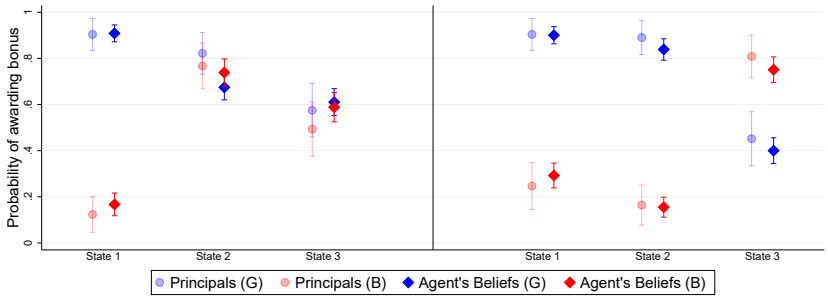
Turning to the reward-after treatment, Figure 4.3 presents some first evidence that agents are capable of anticipating the principals' OB. The figure displays the agents' average belief of obtaining the bonus in the reward-after treatment, conditional on the chosen lottery and the realized states. Overall, the agents' average beliefs follow the patterns of the principals' actual bonus decisions quite well. The beliefs capture the principals' tendency to award bonuses for the lottery that yielded a higher outcome in a given state, and also the decreasing strength of this pattern when moving from lottery pair 1 to 4. This suggests that agents do anticipate the principals' OB.

The average beliefs could mask considerable errors in beliefs at the individual level. To explore this possibility, I calculate, for each agent and choice-state combination, the absolute deviation of the agent's stated belief from the probability that a randomly chosen principal (across all sessions) awards the bonus. As a benchmark, I consider the absolute error in the reward-before treatment. As OB cannot influence bonus decisions in the reward-before treatment, this choice of benchmark seems adequate to assess whether a failure to anticipate OB induces erroneous beliefs. The average absolute error in the reward-after treatment is 18.2 ppts, which is significantly lower than the average absolute error of 22.7ppts displayed by their counterparts in the reward-before treatment ($p < 0.001$).[31] If anything, this suggests that the presence of OB in bonus decisions makes it easier for agents to anticipate the principal's bonus decisions.

[31] The p-values are obtained from two-sided t-tests, clustered at the individual level to control for repeated observations from the same individual.



(a) Lottery pair 1



(b) Lottery pair 4

Table notes: With 95% confidence intervals. For reference, the principals' bonus decisions are displayed in transparent circles.

Figure 4.3 Average beliefs of the 73 agents in the reward-after treatment, conditional on the lottery choice and the realized state of the world.

To understand the potential implications of agents' beliefs for their lottery choices on behalf of the principals, it is important to consider individual beliefs about incentives. A self-interested agent might be expected to choose whichever lottery he believes will give him a higher probability of receiving the bonus. Figure 4.4 displays the proportion of agents who believe to have weakly positive incentives to choose the dominant lottery.[32] For lottery pair 1, a majority of the agents anticipate being strictly more likely to receive the bonus when choosing the dominated lottery. Importantly, this is the case only when the dominated lottery yields a higher payoff in most states and only in the reward-after treatment. For lottery pair 2, a similar pattern is observed (see figure B.4a). This pattern strongly suggest that the agents' beliefs in perverse incentives are driven by an anticipation of the principals' OB. For lottery pair 4 (and also 3, see figure B.4b), a large majority of the agents in both treatments expect positive incentives to choose the dominant lottery under both correlation structures.[33]

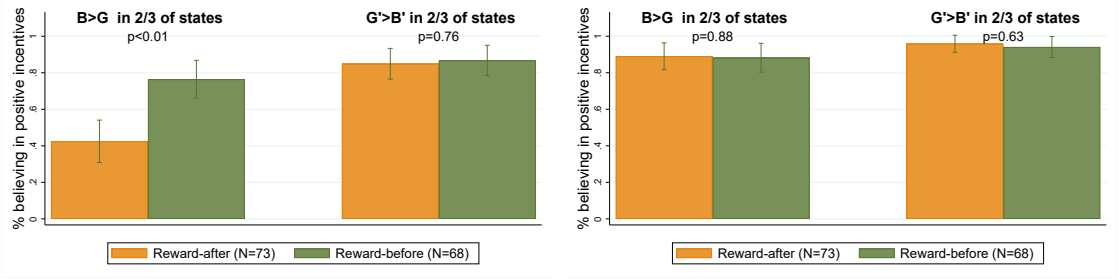
Result 6. ***The agents' beliefs.***

a) In the reward-after condition, the agents' average beliefs are well aligned with the principals' bonus decisions, which indicates that the agents anticipate the principals' OB.

[32] The beliefs about incentives are calculated as follows. Denote $E_a[Pr_p(\theta_k^{corr})]$ the belief of agents a that he will receive the bonus when choosing lottery θ of lottery pair k under correlation structure $corr$ and state s materializes. For the reward-after treatment, the beliefs over the incentives to choose the dominant lottery are calculated as $E_a[I(G_k^{corr})] = \sum_{s\in\{1,2,3\}}(E_a[Pr_p(G_k^{corr}, after, s)] - E_a[Pr_p(B_k^{corr}, after, s)]$. In the reward-before treatment, agents are asked to state their beliefs unconditional of the state, and beliefs over their incentives are calculated accordingly.

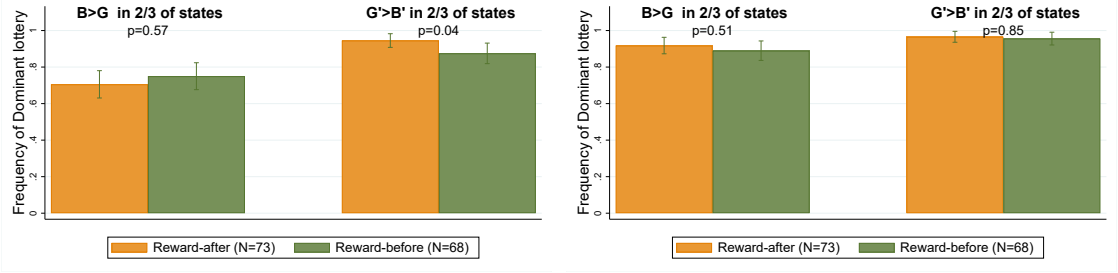
[33] For a detailed analysis, see section B.3.1.4 in the appendix.

b) For lottery pairs 1 and 2 majority of the agents anticipate being more likely to receive the bonus when choosing the dominated lottery. Importantly, this is only true when it is the dominated lottery that yields a higher payoff more often and only in the reward-after treatment.



(a) Lottery pair 1 **(b)** Lottery pair 4

Figure 4.4 The proportion of agents who believe to have weakly positive incentives to choose the dominant lottery, with 95% confidence intervals.



(a) Lottery pair 1 **(b)** Lottery pair 4

Table notes: Choice frequencies are averaged over the first-second choice agents make for each lottery task, with 95% confidence intervals

Figure 4.5 The frequency with which agents choose the dominant lottery.

4.5.5 The agents' lottery choices

The agents' beliefs about their incentives to choose the dominant lottery displayed in figure 4.4 set clear expectations about agents' lottery choices. However, the agents' actual choices display a remarkably different pattern. Consider figure 4.5 that shows the agents' choices averaged over the two choices they make.[34] Overall, agents choose the dominant lottery at high frequencies, and choices are similar across treatments, even for the choice tasks for which their stated beliefs would suggest otherwise.

I now turn to the preregistered hypotheses. Hypothesis 7a states that agents in the reward-before treatment will choose the dominant lottery at high and roughly equal frequencies under both correlation structures for all four lottery pairs. I test this hypothesis by testing the null that the frequencies with which the agents choose the dominant lottery differs across correlation structure for each lottery pair using a two-sided Wilcoxon signed-rank test.[35] Except for lottery pair 3, agents choose the dominant lottery at a significantly higher frequency when it yields a

[34] In averaging the choices over the first and second choices, I follow the preregistration. Aggregate first and second-round choices do not differ significantly for any of the eight choice tasks in either of the two treatments ($p > 0.179$, Exact McNemar test).

[35] I use the Wilcoxon signed-rank test because the basis for the test is the frequency with which an agent chose the dominant lottery, averaged over the two choices he made.

higher outcome in two out of three states ($p < 0.05$, two-sided Wilcoxon signed-rank test).[36] I therefore reject hypothesis 7a.

Hypothesis 7b states that agents in the reward-after treatment are less likely to choose the dominant lotteries than their counterparts in the reward-after treatment under both correlation structures. I test the null hypothesis that $F(G_k^{corr}, before) = F(G_k^{corr}, after)$ against the alternative that $F(G_k^{corr}, before) > F(G_k^{corr}, after)$, using one-sided Wilcoxon rank-sum tests. I reject the null hypothesis at $p < 0.05$ in favor of the alternative only for one of the eight comparisons, namely for lottery 2, when the dominant lottery yields a lower payoff in only one state. For this choice task, agents in the reward-after treatment choose the dominant lottery at a frequency of 68.5% whereas their counterparts in the reward-before treatment chose it at 81.6% frequency.

Hypothesis 7c states that changing the correlation structure from the "one-state-better" to the "two-state-better" correlation structure increases the frequency of dominant lottery choices more in the reward-after than in the reward-before treatment. Formally, I test the null that $F(G_k^{1s}, before) - F(G_k^{2s}, before) = F(G_k^{1s}, after) - F(G_k^{2s}, after)$ against the alternative that $F(G_k^{1s}, before) - F(G_k^{2s}, before) > F(G_k^{1s}, after) - F(G_k^{2s}, after)$. I reject the null hypothesis at $p = 0.043$ for lottery pair 1 and at $p = 0.027$ for lottery pair 2, and do not reject it for lottery pairs 3 and 4 at any conventional level. However, given that I reject hypothesis 7b for lottery pair 1, the rejection of the null for lottery pair 1 seems to be partially driven by the fact that agents in the reward-after treatment choose the dominant lottery G_k^{2s} somewhat more often than their counterparts in the reward-before treatment. Therefore, there is support for the notion that changing the correlation structure from "one-state-better" to "two-state-better" increases the frequency of dominant lottery choices due to agents anticipating the principals' OB only for lottery pair 2.

Overall, the evidence that anticipated OB induces agents to choose the principal's least preferred option is weak at best. I summarize the results on the agents' lottery choices below.

Result 7. ***The agents' lottery choices.***

a) Agents in the reward-before condition choose the dominant lottery at a significantly higher frequency in the reward-before than in the reward-after treatment. I reject hypothesis 7a.

b) Agents in the reward-after treatment choose the dominant lottery significantly less often than the agents in the reward-before treatment only for lottery pair 2 when the dominant lottery yields a higher payoff in only one state. I reject hypothesis 7b.

c) I confirm hypothesis 7c for lottery pairs 1 and 2 but not for lottery pairs 3 and 4. However, for lottery pair 1, the rejection of the null is caused by a tendency of agents in the reward-after treatment to choose the dominant lottery G_k^{2s} more often than their counterparts in the reward-before treatment.

[36] Table B.4 provides an overview of the choice frequencies of the the dominant lotteries and relevant test statistics.

4.5.6 Discussion of the agents' choices

The preceding analysis at the aggregate level suggests that lottery choices are largely inconsistent with stated beliefs, but only when the stated beliefs of a majority imply a choice of the dominated lottery. Individual-level analysis, relegated to the appendix (section B.3.2) confirms this impression. It cannot be entirely ruled out that agents might disregard their monetary incentives because of a moral obligation to make to make choices that are in line with their principals' preferences as suggested by Kling et al. (2023). However, in appendix AB.3.2, I show that models of altruism or inequity aversion (Fehr and Schmidt, 1999) cannot convincingly rationalize the agents' behavior.[37]

Exploratory analysis suggests that agents' behavior can be explained at least partially by a failure of strategic reasoning. Building on an argument made by Costa-Gomes and Weizsäcker (2008), stated beliefs might reveal a deeper strategic understanding than actions, and actions might not necessarily result from best-responding to beliefs.[38] Several pieces of evidence point to this explanation. First, response times suggest that many agents might not form beliefs at all before their initial choice. Recall that agents made a first choice before any belief elicitation, and a second choice after the belief elicitation. Averaging over all choice tasks, agents in the reward-after treatment take an average of 15.3 seconds to make their first choice. Had they already formed beliefs at this stage, entering them subsequently should be quick. However, the average agent spends 55.1 seconds to enter his beliefs for a given lottery task. Moreover, forming beliefs about incentives is arguably easier in the reward-before treatment, because agents need not think through all possible states. However, agents in the reward-before treatment take an average of 14.9 seconds for their first choice, which does not differ significantly from the 15.3 seconds of their counterparts in the reward-after treatment ($p = 0.760$, two-sided clustered t-test). Both pieces of evidence are consistent with the conjecture that many agents simply do not form beliefs before being prompted to do so.

Second, (strategically) sophisticated agents tend to align their choices more with their beliefs, but only after they have been prompted to form beliefs. As proxies for agents' strategic sophistication, I consider their CRT-score and whether they major economics.[39] I focus the analysis on lottery pairs 1 and 2 under the 1-state-better correlation structure, as these are the choice tasks for which choices and beliefs are largely inconsistent. In the following, I further focus on the choices individuals make immediately after the belief elicitation. Results for first-round choices can be found in the appendix (Table B.6). In table 4.6a, I estimate random-effects panel models. The dependent variable is the dummy *consistent*, which indicates whether a given choice is consistent

[37] For altruism to rationalize the agents' behavior, many of agents would have to assign more weight to the principals' payoff than to their own. Holding the probability of receiving a bonus constant, inequity averse agents would prefer the principal to obtain the dominated lottery, as this makes them less behind the principal in case a medium or high payoff realizes.

[38] Since the agents' stated beliefs align well with the principals' actual bonus decision, it seems unlikely that they do not reflect the agents' true beliefs.

[39] The CRT-score is a well-established measure of a tendency to engage in cognitive reflection and correlated well with other measures of cognitive ability (Frederick, 2005; Toplak et al., 2014). All economics majors should have had some exposure to game theory by the time the experiment was run, which might arguably induce more strategic thinking.

with stated beliefs.[40] In column (1), the explanatory variables are a dummy indicating whether an agent's beliefs indicate weakly positive incentives to choose the dominant lottery (*positive*), their CRT score, and an interaction between the two. The coefficient on the dummy *positive* indicates that experimental subjects who believe in positive incentives are 59.7 ppts ($p < 0.001$) more likely to satisfy consistency. Further, the estimated coefficient on the CRT-score suggests that, among subjects who believe they have strict incentives to choose the dominated lottery, a one-point increase in the CRT-score is associated with a 7.4ppt ($p = 0.006$) increase in the likelihood of satisfying consistency. The coefficient on the interaction term is estimated at -0.038 and is statistically insignificant ($p = 0.296$). A chi-square test of the additive effect of the CRT-score and the interaction term suggests that the CRT-score does not significantly predict increased consistency when agents have positive beliefs ($p = 0.174$).

Figure 4.6b illustrates the relationship between the CRT score and choosing the dominant lottery for individuals who believe to have incentives to choose the dominated lottery. The frequency of dominant lottery choices decreases, almost monotonically, from 100% for a CRT score of 0 to 29% for individuals scoring the maximum 7 points. These results suggest that principals high in cognitive reflection do choose the dominated lottery at elevated rates when they believe to have an incentive to do so.

Column (2) presents results from a similar exercise, but replaces the CRT score with a dummy indicating whether a subject majors in economics. Again, consistency is significantly higher when agents have positive beliefs. Further, when agents have negative beliefs, they are 30.2ppts more likely to be consistent when they major in economics ($p = 0.006$). When agents have positive beliefs, studying economics does not significantly predict consistency.[41]

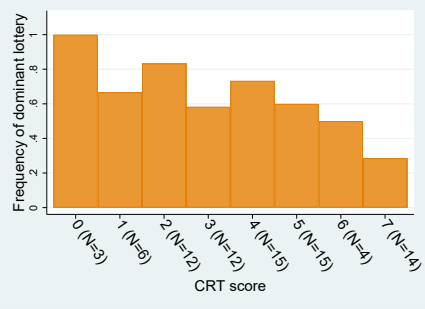
If the CRT score and studying economics predict greater consistency because the variables proxy for a greater propensity to engage in strategic reasoning, they should not be predictive of consistency in the reward-before treatment where beliefs almost always imply a choice of the dominant lottery. In columns (3) and (4), the regression exercises are therefore repeated for agents in the reward-before treatment. Reassuringly, neither the CRT score, nor whether subjects study economics significantly predicts consistency.

[40] I define a lottery choice to be consistent with an agent's beliefs if an agent believes to have weak incentives to choose the dominant lottery and chooses the dominant lottery or believes to have strictly positive incentives to choose the dominated lottery and chooses the dominated lottery. Otherwise, a choice is said to be inconsistent.

[41] A chi-square test suggests that economics students are no more consistent than participants from other fields of study when they hold positive beliefs ($p = 0.699$).

	Reward-after		Reward-before	
	(1)	(2)	(3)	(4)
VARIABLES	Consistent	Consistent	Consistent	Consistent
positive	0.597***	0.567***	0.214*	0.136
	(0.147)	(0.110)	(0.117)	(0.083)
CRT	0.074***		0.024	
	(0.027)		(0.031)	
CRT ∗ *positive*	-0.038		-0.014	
	(0.036)		(0.034)	
Econ		0.302***		-0.019
		(0.110)		(0.115)
Econ ∗ *pos*		-0.255*		0.105
		(0.153)		(0.128)
Constant	0.080	0.203**	0.686***	0.760***
	(0.119)	(0.082)	(0.103)	(0.073)
Observations	146	146	136	136
Number of subjects	73	73	68	68

(a) The dependent variable consistent is equal to 1 if an agent's stated beliefs indicate weakly positive incentives to choose the dominant lottery and the agent chooses the dominant lottery or if beliefs indicate strictly positive incentives to choose the dominated lottery and the agent chooses the dominated lottery. Otherwise, the variable equals 0. The sample consists of second choices agents make after the belief elicitation. The table displays coefficients of random-effects regressions. Standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.



(b) Relationship between CRT score and choosing the dominant lottery, for individuals who believe to have incentives to choose the dominated lottery. The sample consists of second choices agents make after the belief elicitation.

In summary, it appears that agents fail to form beliefs when asked to make their initial choices. Once being prompted to form beliefs, sophisticated agents seem to understand that they might be better off choosing the dominated lottery and show a higher propensity to do so when they believe that this in their best interest. These findings suggest that a certain level of strategic sophistication are a prerequisite for anticipated OB to translate into actions that are harmful for principals.

4.6 Structural Estimation

In this section, I provide an estimation of the principals' degree of OB and the agent's anticipated degree of OB, both at the individual and the aggregate level.

4.6.1 Estimation framework

I estimate the principals' degree of OB using a random utility approach. I lay out the main ideas of the approach here and provide a detailed treatment in appendix B.4. Assuming a type-1 extreme value distribution of errors, the probability with which a principal awards the bonus if an agent chooses lottery θ_k^{corr} and state s materializes follows the convenient logit form (McFadden, 1973).

$$pr(\lambda_p, \mu_p^1, \theta_k^{corr}, s) = \frac{1}{1 + exp(-\mu_p^1 z(\theta_k^{corr}, s))} \tag{4.4}$$

,where μ_p^1 is a noise parameter that captures how well the posited function explains a principal's bonus decisions and

$$z(\theta_k^{corr}, s) = \lambda_p [x_s^{\theta_k^{corr}} - x_s^{-\theta_k^{corr}}] + (1 - \lambda_p)[EV(\theta_k^{corr}) - EV(-\theta_k^{corr})] \tag{4.5}$$

I impose a linear utility function, that is $u(x) = x$, which implies that the expected utility is

given by the expected value, $EU(\theta_k^{corr}) = \sum_s q_s x_s^{\theta_k^{corr}}$. In appendix AB.5.1, I allow for $u(\cdot)$ to be described by a CRRA utility function. Likelihood ratio tests suggest that imposing risk neutrality does not necessarily harm the model fit, and the estimated degrees of OB stay virtually unchanged. I estimate the parameters (λ_p, μ_p^1) via maximum likelihood estimation from 54 bonus decisions.

Agents are asked to state their belief about the probability with which they will receive a bonus when matched to a random principal for all 54 possible lottery choice and state combination. I assume that agents have in their mind the logit formulation given in equation 4.4 and form beliefs about the relevant parameters. In order to keep the estimation tractable I impose that agents imagine a representative principal. That is, they have to form expectations over λ, and μ^1. I denote the beliefs of agent a over these parameters by $\check{\lambda}_a$, and $\check{\mu}_a^1$. I impose that agents implement their beliefs with a mean-zero error that is normally distributed with variance given by $1/\mu_a^2$. Hence, μ_a^2 can be interpreted as a noise parameter governing how precisely agents state beliefs. Using maximum likelihood estimation, I estimate, for each agent, the three parameters $(\check{\lambda}_a, \check{\mu}_a^1, \mu_a^2)$.

For the estimation, I exclude 11 principals and six agents because their data cannot be used to obtain reliable estimates.[42] The exclusion rate of 11.6% is comparable to, if not slightly lower than, that observed in many studies employing structural estimation.[43] Including these subjects does not change results qualitatively.

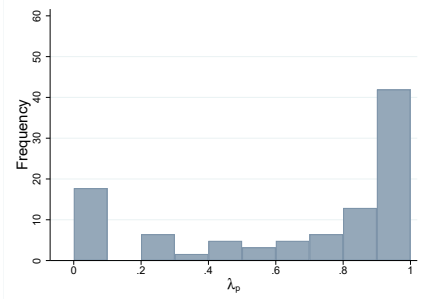
4.6.2 Results - Principals

4.6.2.1 Individual level estimates

Figure 4.7a present individual-level estimates of λ_p. As can be seen, the estimates display considerable heterogeneity. The distribution of λ_p appears to be approximately bimodal, with 17.7% of the principals having an estimated $\lambda_p < 0.1$ and 41.9% having an estimated $\lambda_p > 0.9$. This suggests that most principals make their bonus decisions following one decision rule and reward based on choices or based on outcomes but not some combination of both. Manual inspection of the bonus decisions confirms this impression. A number of principals implement one of the rules: "reward iff the agent chose the dominant lottery" or "reward iff the agent obtained a higher outcome" perfectly or near perfectly.

[42] 13 subjects are excluded because their choices are too noisy for the estimation. Four principals (almost) always give the bonus to the matched rather than the randomly chosen agent. These subjects follow a precise strategy, but their data cannot be used to estimate their degree of OB. I identify the subjects I exclude by estimating the model for each subject and then visually inspecting the choices of subjects with very low noise parameters.

[43] Exclusion rates of comparable studies are 17.6% in Van Leeuwen and Alger (2019), 26.3% in Fisman et al. (2007), 19.5% in Bleichrodt et al. (2010), and 8% in Bruhin et al. (2019).



(a) Individual level estimates of the OB parameter λ_p from 62 principals in the reward-after treatment.

	(1) λ_p	(2) λ_p	(3) λ_p	(4) λ_p	(5) λ_p
CRT score	-0.065***				
	(0.021)				
	[0.012]				
above median seconds		-0.376***			
		(0.083)			
		[0.003]			
"rewarded choices"			0.004		
			(0.022)		
			[0.877]		
"rewarded outcome"				0.066***	
				(0.017)	
				[0.009]	
"rewarded counterfactual"					0.064***
					(0.016)
					[0.009]
Constant	0.860***	0.856***	0.642***	0.256**	0.276**
	(0.070)	(0.039)	(0.160)	(0.127)	(0.118)
Observations	62	62	62	62	62
R-squared	0.127	0.255	0.000	0.217	0.244

(b) OLS regression relating individual-level estimates of the OB parameter λ_p to individual characteristics.

Table notes: CRT score ranges from 0-7, with higher values indicating higher cognitive reflection. The "above median seconds" equals 1 if a subject's average response time on bonus decisions is above the median. The variables "rewarded ..." indicate to which extent subjects based bonus decisions on the agent's choices, their outcome, or counterfactual comparison between the obtained and the forgone outcome, based on self-reports on a scale from 1-9. Robust standard errors are in parentheses. The sample consists of 62 principals in the reward-after condition. MHT corrected Romano-Wolf step-down p-values in square brackets. *** p<0.01, ** p<0.05, * p<0.1, referring to the MHT corrected p-values (except for the constants).

A key advantage of the structural estimation is that it allows exploring associations between individual characteristics and the obtained estimates. This helps shed light on the underlying psychological mechanisms. Table 4.7b reports results from a number of OLS regressions. The regression exercises reported in columns (1) and (2) suggest that OB might indeed be caused by a cognitive channel, as assumed in the model. Column (1) regresses λ_p on the participants' CRT score. The coefficient on the CRT score is estimated at -0.065, which suggests that individuals higher in cognitive reflection display less OB. Correcting for multiple hypothesis testing (MHT) (Romano and Wolf, 2005, 2016), the coefficient is significant at $p = 0.012$. Given that the CRT-score ranges from 0-7, the magnitude of the effects is considerable. Column (2) regresses λ_p on a dummy indicating whether a participant's response time, averaged over all bonus decisions, is above the median response time. The estimate suggests that individuals with an above-median response time have a 0.376 points lower degree of OB ($p = 0.003$, corrected for MHT). The association between longer response times and lower levels of OB can be seen as further evidence suggesting that OB arises from a cognitive channel. The findings corroborate the view that the misalignment between the principals' ex-ante preferences and their bonus decisions results from a mistake they make when conditioning bonus decisions on outcomes.

Columns (3), (4), and (5) provide evidence that self-reported strategies are predictive of the estimates of OB. Participants were asked to indicate, on a Likert scale from 1-9, to which extent they rewarded based on choices ("rewarded choices", column 3), the realized outcome ("rewarded outcome", column 4), or based on a counterfactual comparison of the realized and the forgone outcome ("rewarded counterfactual", column 5). Whereas the estimated coefficient on "rewarded choices" is close to zero and statistically insignificant, the coefficients on "rewarded outcome" and

"rewarded counterfactual" are estimated at 0.066 and 0.064 and are both statistically significant at $p = 0.009$ (corrected for MHT). As the variables range from 1-9, the magnitudes of these effects is sizable. The finding that self-reported strategies predict the estimated degree of OB lends credibility to the estimates.[44]

One further aspect to note about the regression exercise is that the explanatory variables, with the exception of "rewarded choice", explain a considerable part of the variation in the estimated degree of OB. The r-squares range from 0.127 for the CRT-score to 0.255 for response time.[45] This suggests that the estimated parameters of OB capture a meaningful aspect of individual behavior, which lends additional support not just to the structural estimation, but also to the underlying model of OB.

4.6.2.2 Aggregate level estimation

I provide aggregate level estimates in table 4.6. Column (1) displays estimates of the representative principal. She is estimated to have a degree of OB equal to $\lambda = 0.589$. The representative principal thus bases bonus decision roughly 60% on the ex-post comparison of outcomes and 40% on the agents' choices. The estimates square well with the individual-level estimates that suggest high levels of OB. The noise parameter is estimated at $\mu^1 = 4.524$. The estimate implies that choice probabilities are well distinguished from the 50% random choice benchmark, but are also far from deterministic. Following the individual-level estimates that suggest that most principals either display very high or very low levels of OB, column (2) presents a specification that assumes that principals are either fully outcome biased, or fully unbiased and estimates the proportion of each category. I impose a common noise parameter for these two types. The fraction of fully outcome biased principals is estimated to be 65.9%, and the estimate of the noise parameter is equal to 6.298. Finally, column (3) allows for two separate noise parameters for the fully and the unbiased type. The Fraction of fully outcome biased principals is now estimated at 80.3%. The noise parameter for the fully outcome biased type is estimated at 3.638, whereas that of the unbiased type is estimated at 31.961. These estimates suggest that the unbiased type captures a small and homogeneous group of unbiased principals who reward only based on choices and implement their bonus decisions with very little error. The full OB type seems to capture a larger group of principals who reward based on outcomes but do so more nosily.

The Akaike and the Bayesian information criteria (AIC, and BIC) both suggest that the model in column (3) yields the best fit, followed by that in column (2), while the representative principal model comes in last.[46] This confirms the impression gained from the individual level analysis that most principals display either a very high or a very low level of OB. In appendix B.5.2, I discuss results from the estimation of finite mixture models that further confirm this expression. Overall,

[44] A possible explanation for why "rewarded choices" does not correlated with λ_p is that most subjects indicated very high values. The variance of "rewarded choices" is somewhat smaller than that of the other two variables, but only significantly so for "rewarded counterfactual" ($p = 0.125$ "rewarded choices" and "rewarded counterfactual" $p = 0.028$, two-sided variance ration test).

[45] By way of comparison, Dohmen et al. (2011) obtain an r-squared of 0.12 when regressing the self-reported willingness of risk, which is generally thought to correlated well with relevant measures, on a dummy indicating gender, height, and age.

[46] A Likelihood ratio test between the model in column (2) and (3) confirms the impression that imposing the restriction $\mu^1(\lambda = 0) = \mu^1(\lambda = 1)$ decreases the fit of the model significantly ($\chi^2(1) = 33.904$, $p < 0.001$).

the results suggest that heterogeneity in OB can be modelled parsimoniously by assuming a fully outcome biased and a fully unbiased type, a result that might be of interest for theorists.

	(1) Representative Principal	(2) Full or no OB	(3)
λ	0.589		
	(0.085)		
Fraction $\lambda = 1$		0.659	0.803
		(0.070)	(0.048)
μ^1	4.524	6.298	
	(0.417)	(1.597)	
$\mu^1(\lambda = 0)$			31.961
			(7.298)
$\mu^1(\lambda = 1)$			3.638
			(0.413)
Individuals	62	62	62
Observations	3348	3348	3348
$logL$	-1648.012	- 1637.971	- 1621.019
BIC	3304.278	3284.197	3254.420
AIC	3300.023	3279.942	3248.038

Table notes: Column (1) assumes a representative principal. Columns (2) and (3) assume that there is one type with $\lambda = 1$ and another type with $\lambda = 0$, and estimate the fraction of subjects in each category. Column (2) imposes a common noise parameter, whereas column (3) allows for two different noise parameters. Standard errors, in parentheses, are obtained from cluster-bootstrap with 4000 repetitions.

Table 4.6 Aggregate level estimates of the principals' OB.

4.6.3 Results - agents

4.6.3.1 Individual level

Figure 4.8 displays individual level estimates of the anticipated level of OB. The estimates assume that agents imagine a representative principal when forming their beliefs, as this ensure better comparability with the individual levels estimates on the principals' side. Imposing that agents imagine that principals are either fully outcome biased or fully unbiased produces similar results. See appendix B.5.3 for details. 52.2% of the estimated $\check{\lambda}_a$ are larger than 0.9, which suggests that about half of the agents expect the representative principal to be fully or near fully outcome-biased. The remaining 47.7% of the estimated $\check{\lambda}_a$ seem approximately uniformly distributed between 0 and 0.9. Importantly, there is no concentration at or close to zero, as might be expected if agents project their own degree of outcome bias on the principals. Overall, the individual level estimation seems to suggest that, although the beliefs of the average agent track the broad patterns of the principals' bonus decisions well, a number of them overestimates the principals' degree of OB. Moreover, there does not seem to be as much heterogeneity in anticipated than in actual OB. In appendix B.5.3, I discuss regression analysis similar to that conducted for the principals' degree of OB, which confirms the impression of limited heterogeneity in anticipated OB.

4.6.3.2 Aggregate level estimates

Table 4.7 displays aggregate level estimates for the anticipated degree of OB. The specifications mirror those estimated on the principals' side. Column (1) assumes that agents imagine a repre-

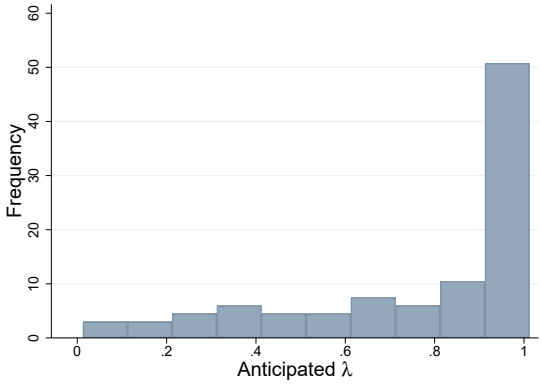


Table notes: The displayed estimates are of 67 agents in the reward-after treatment. The estimates impose that agents imagine a representative principal when forming beliefs.

Figure 4.8 Individual level estimates of the agents' beliefs $\check{\lambda}$.

sentative principal. The Representative agent is estimated to anticipate a representative principal with a degree of OB by 0.788, which suggest a high level of anticipated OB. Column (2) imposes that agents imagine that principals are either fully outcome biased or fully unbiased, and estimates the anticipated fraction of each type. For this specification, I impose a common anticipated noise parameter $\check{\mu}^1$ for both types. The representative agent anticipates that 78.4% of the principals are fully outcome biased. Finally, column (3) allows for two separate anticipated noise parameters for the full OB and the no-OB type. The estimate of the anticipated fraction of fully outcome biased principal is increased to 0.891. The anticipated noise parameter for the no-OB type is estimated at a very high 72.290, whereas the anticipated noise parameter of the full OB type is estimated at 3.465. This suggests that the representative agent expects the no-OB type to implement bonus decisions almost without error, which mirrors the estimates on the principals' side.

An advantage of the structural estimation is that it allows for formal tests of whether agents anticipate the correct degree of outcome bias. All estimates of anticipated OB are somewhat higher than their counterparts of the estimated principals' degree of OB. Using two-sided Wald tests, these differences are statistically significant at $p = 0.058$ for the column (1) specification, at $p = 0.186$ for the column (2) specification, and at $p = 0.118$ for the column (3) specification. Overall, this confirms the impression that the representative agent's anticipated degree of OB is well calibrated. If anything, he slightly overestimates but certainly does not underestimate the principals' degree of OB.

4.6.4 Robustness

I conduct a number of additional estimation exercises to ensure the robustness of the results. Details can be found in appendix B.5.4. There are mainly two issues that could lead to biased estimates of the principals' degree of OB and the agents' anticipation thereof. First, as λ loads on the difference between the obtained and the forgone payoff, the estimates discussed thus far might partially reflect a tendency to reward high obtained outcomes, irrespective of a comparison to the forgone payoff. A second concern is correlation-sensitivity in ex-ante preferences. Fortunately,

	(1)	(2)	(3)
	Representative Principal	Full or no OB	
$\check{\lambda}$	0.788		
	(0.064)		
Fraction $\lambda = 1$		0.784	0.891
		(0.051)	(0.029)
$\check{\mu}^1$	3.872	4.690	
	(0.323)	(1.417)	
$\check{\mu}^1(\lambda = 0)$			72.290
			(16.623)
$\check{\mu}^1(\lambda = 1)$			3.631
			(0.355)
μ^2	3.439	3.457	3.465
	(0.167)	(0.171)	(0.163)
Individuals	67	67	67
Observations	3618	3618	3618
$logL$	-1254.479	- 1244.154	- 1226.435
BIC	2517.367	2496.717	2465.484
AIC	2512.958	2492.308	2458.87

Table notes: Column (1) assumes that agents imagine a representative principal. Columns (2) and (3) assume that agents imagine that there is one type of principals with $\lambda = 1$ and another type with $\lambda = 0$, and estimate the fraction of subjects in each category. Column (2) imposes a common noise parameter, whereas column (3) allows for two different noise parameters. Standard errors, in parentheses, are obtained from cluster-bootstrap with 4000 repetitions.

Table 4.7 Aggregate level estimates of the agent's anticipated degree of OB.

both concerns can be addressed by adjusting the estimated functions by allowing bonus decisions to depend on the obtained and forgone outcome independently and by allowing for correlation-sensitive preferences. Results remain qualitatively similar. The individual level results remain virtually unchanged. Aggregate estimates of the degree of OB and the anticipated degree of OB are somewhat reduced, but do not change significantly in most cases.

4.7 Conclusion

In this paper, I investigate whether a tendency to reward economic agents as if they could have anticipated a random outcome can induce agents to take sub-optimal actions. I provide a simple model of OB that suggests that OB might trick principals into incentivizing agents to choose their least preferred action. I provide experimental evidence consistent with this prediction. According to stated beliefs, agents anticipate the principals' outcome bias, and a majority expects to be more likely to obtain a fixed bonus when choosing a dominated lottery, when this lottery is more likely to yield a higher outcome than the dominant alternative. However, in what appears to be a failure of strategic reasoning, most agents choose the dominant lottery nevertheless. Experimental subjects high in cognitive reflection and economic students are more likely to choose the dominated lottery when they believe to have an incentive to do so.

Structural estimation provides additional insights into the nature of outcome bias. The distribution of OB appears to be nearly bimodal, with most subjects displaying either (nearly) full or (nearly) no OB at all. Further, individuals' degree of OB is negatively associated with a measure of cognitive reflection and response time, which suggests that the OB is driven by a cognitive channel.

Overall, the results suggest that OB might be most detrimental in settings in which relatively unsophisticated principals meet sophisticated agents. This description seems to fit many settings of delegated expertise. Sophisticated individuals are likely to select themselves into important positions in which they act as agents, may that be in politics, corporate governance, or investment management. At the same time, principals, such as voters or ordinary citizens seeking a mutual fund to invest for retirement might often lack sophistication, especially, in the agent's domain of expertise.

Chapter 5

Conclusion

In conclusion, my Ph.D. thesis investigates whether and in which situations decision-making under risk is correlation-sensitive through a combination of theory and controlled lab experiments. Chapters 1 and 2 follow the tradition of decision-theoretical research and study individual decision-making, whereas Chapter 3 ventures outside of the realm of individual decision-making and studies correlation sensitivity in a setting of delegated decision-making.

Overall, what should the reader take away from this thesis? The main result of Chapter 1 is that most participants of our experiments are not correlation-sensitive and that only a minority displays consistent correlation sensitivity. Moreover, this minority displays DSPD. From the viewpoint of regret and salience theory, this is the "wrong" kind of correlation sensitivity. "Wrong" in the sense that it cannot explain the deviations from the expected utility theory economists are usually interested in. Chapter 2 provides evidence that recent experimental findings that seemed to show evidence in favor of ISPD were most likely caused by confounds in the experimental design, not correlation sensitivity. Given that it is ISPD that allows correlation-sensitive theories of decision-making under risk to rationalize important deviations of EUT, our results suggest that correlation-sensitivity might not be the key to understanding these deviations. Given the complexities implied by theories that violate transitivity, a possible conclusion is that economists might be better off discarding correlation sensitivity altogether and spending their time more fruitfully on theories that do satisfy transitivity.

I would caution against this conclusion. First, in my personal experience, especially proponents of regret and salience might not accept the conclusion that there is little evidence for ISPD and that the preferences of correlation-sensitive individuals predominantly satisfy DSPD. One reason for this reluctance is that one method of eliciting preferences, the trade-off method, has consistently produced evidence favoring a predominance of ISPD (Bleichrodt et al., 2010; Baillon et al., 2015). An important avenue for future research is understanding why different experimental methods yield different results. This might help in achieving a consensus on the issue of the empirical content of correlation-sensitive preferences.

Second, I would argue that correlation-sensitive preferences are an interesting phenomenon in their own regard, even if they do not seem to be the cause for the most commonly studied

deviations from EUT. More research is needed to better understand them. In Chapter 1, we start investigating why people might display correlation sensitivity and find that deliberate within-state payoff comparisons play an important role. However, it is not clear why this is the case. Given that we do not have a good understanding of the nature of correlation-sensitive preferences at this point, it seems conceivable that they do play an important role in some settings. A better understanding of the causes of correlation sensitivity might help gauge their importance in decision-making under risk.

Chapter 3 suggests that correlation sensitivity might be a fruitful lens through which to understand the choices of agents who are subjected to an outcome-biased principal. I find that experimental subjects in the role of principal have a strong tendency to reward agents if they choose an action that happens to yield a higher payoff than the available alternative. This can induce correlation sensitivity in the agents' incentives, although the principals' own choices do not show evidence of correlation-sensitive preferences. Agents tend to anticipate the principals' outcome bias. However, I find that only strategically sophisticated agents also display correlation sensitivity in their choices. Structural estimation suggests that there is stark heterogeneity in the principals' outcome bias and that high outcome bias is strongly associated with low cognitive reflections. These results suggest that agents' choices most likely display correlation sensitivity where high-sophistication agents deal with low-sophistication principles.

Chapter 3 opens up many avenues for future research. Below, I sketch out just two of them. The first interesting direction is learning. Do principals learn to be less outcome-biased over time? Do agents learn to behave in a way that maximizes monetary payments? Will learning and experience increase or reduce the outcome bias of principals and the correlation-sensitivity displayed by agents? How does this depend on the specifics of the interactions? In some settings, the same agent (an expert) might interact with many different principals (one-time customers), and learning might occur predominantly on the agent's side. In this setting, it seems difficult to imagine that agents do not learn how to exploit their principal's bias. In other settings, learning might predominantly occur on the principal's side. A priori, this kind of setting seems more conducive to principals conquering their outcome bias. In other settings, interactions might be repeated, with learning opportunities for both agents and principals.

A second interesting question is about how to mitigate potentially harmful outcome bias. Can we put specific procedures or nudges into place for this purpose? Are there effective strategies agents can employ by their own initiative to insure themselves against their principal's outcome bias without sacrificing efficiency? Would markets help in mitigating problems arising from outcome bias? Much might depend on the specificities of the market. On the one hand, competition might force highly outcome-biased principals to exit the market. On the other hand, if a market mechanism selects agents who are more able or willing to exploit their principals' outcome bias into positions where they can profitably do so, markets might even exacerbate the consequences of outcome bias.

Finally, I want to draw an important conceptual distinction regarding how correlation sensitivity emerges in the different settings considered in this thesis. In Chapters 1 and 2, experimental

subjects are informed about the joint distribution of different lotteries and are asked to choose between them. Therefore, correlation sensitivity might arise because subjects compare payoffs within-state in a *forward-looking* way. In Chapter 3, principals need not be forward-looking to create correlation-sensitive incentives. On the contrary, the within-state comparisons they engage in are inherently *backward-looking*. They react to rather than anticipate the within-state comparisons.

I believe that this distinction points the way to yet another interesting direction for research on correlation sensitivity. Pilot data from an ongoing project (not reported here) suggests that also individual decision-making under risk can become very strongly correlation-sensitive because people compare payoffs within-state in a backward-looking way. Subjects had to make choices analogous to the one displayed in Table 1.1. As in the chapter 1 and 2, they were fully informed about the joint distribution of payoffs. However, they had to make decisions repeatedly and received outcome feedback after each choice. While participants display only very little correlation sensitivity initially, most participants display correlation sensitivity after a while.

If the suspicion that backward-looking correlation sensitivity is much more prevalent than its forward-looking counterpart is correct, correlation sensitivity could be important in many settings. These include individual decision-making, learning, delegated decision-making, and strategic interactions more broadly. A lot of exciting research lies ahead!

Bibliography

Achen, C. H. and L. M. Bartels (2017). Democracy for realists. In *Democracy for Realists*. Princeton University Press.

Aimone, J. A. and X. Pan (2020). Blameable and imperfect: A study of risk-taking and accountability. *Journal of Economic Behavior & Organization 172*, 196–216.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine. *Econometrica 21*(4), 503–546.

Alós-Ferrer, C. and A. Ritschel (2022). Attention and salience in preference reversals. *Experimental Economics*, 1–28.

Andersson, H., H. Sholtz, and J. Zheng (2023). Measuring regret theory in the health and financial domain. *Working Paper*.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 399–405.

Apesteguia, J. and M. A. Ballester (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy 126*(1), 74–106.

Baillon, A., H. Bleichrodt, and A. Cillo (2015). A tailor-made test of intransitive choice. *Operations Research 63*(1), 198–211.

Baron, J. and J. C. Hershey (1988). Outcome bias in decision evaluation. *Journal of personality and social psychology 54*(4), 569.

Becker, G. S. (1976). Altruism, egoism, and genetic fitness: Economics and sociobiology. *Journal of economic Literature 14*(3), 817–826.

Bell, D. E. (1982). Regret in decision making under uncertainty. *Operations research 30*(5), 961–981.

Bell, D. E. (1983). Risk premiums for decision regret. *Management Science 29*(10), 1156–1166.

Bernheim, B. D. (2021). In defense of behavioral welfare economics. *Journal of Economic Methodology 28*(4), 385–400.

Bernheim, B. D. and A. Rangel (2009). Beyond revealed preference: choice-theoretic foundations for behavioral welfare economics. *The Quarterly Journal of Economics 124*(1), 51–104.

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica 22*(1), 23–36.

Bertrand, M. and S. Mullainathan (2001). Are ceos rewarded for luck? the ones without principals are. *The Quarterly Journal of Economics 116*(3), 901–932.

Biais, B. and M. Weber (2009). Hindsight bias, risk perception, and investment performance. *Management Science 55*(6), 1018–1029.

Bialek, M. and G. Pennycook (2018). The cognitive reflection test is robust to multiple exposures. *Behavior research methods 50*, 1953–1959.

Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence 22*(7), 719–725.

Bleichrodt, H., A. Cillo, and E. Diecidue (2010). A quantitative measurement of regret theory. *Management Science 56*(1), 161–175.

Bleichrodt, H. and P. P. Wakker (2015). Regret theory: A bold alternative to the alternatives. *The Economic Journal 125*(583), 493–532.

Boles, T. L. and D. M. Messick (1995). A reverse outcome bias: The influence of multiple reference points on the evaluation of outcomes and decisions. *Organizational Behavior and Human Decision Processes 61*(3), 262–275.

Bolton, G. E. and A. Ockenfels (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review 91*(1), 166–193.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012). Salience theory of choice under risk. *The Quarterly journal of economics 127*(3), 1243–1285.

Bordalo, P., N. Gennaioli, and A. Shleifer (2013). Salience and asset prices. *American Economic Review 103*(3), 623–28.

Bordalo, P., N. Gennaioli, and A. Shleifer (2022). Salience. *Annual Review of Economics 14*, 521–544.

Braun, M. and A. Muermann (2004). The impact of regret on the demand for insurance. *Journal of Risk and Insurance 71*(4), 737–767.

Brownback, A. and M. A. Kuhn (2019). Understanding outcome bias. *Games and Economic Behavior 117*, 342–360.

Bruhin, A., E. Fehr, and D. Schunk (2019). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association 17*(4), 1025–1069.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010a). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica 78*(4), 1375–1412.

Bruhin, A., H. Fehr-Duda, and T. Epper (2010b). Risk and rationality: Uncovering heterogeneity in probability distortion. *Econometrica 78*(4), 1375–1412.

Bruhin, A., M. Manai, and L. Santos-Pinto (2022). Risk and rationality: the relative importance of probability weighting and choice set dependence. *Journal of Risk and Uncertainty 65*(2), 139–184.

Camerer, C. (2011). The promise and success of lab-field generalizability in experimental economics: A critical reply to levitt and list. *Available at SSRN 1977749*.

Camerer, C., G. Loewenstein, and M. Weber (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of political Economy 97*(5), 1232–1254.

Camille, N., G. Coricelli, J. Sallet, P. Pradat-Diehl, J.-R. Duhamel, and A. Sirigu (2004). The involvement of the orbitofrontal cortex in the experience of regret. *Science 304*(5674), 1167–1170.

Celeux, G. and G. Soromenho (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of classification 13*, 195–212.

Chan, N. W. and L. Wolk (2023). Reciprocity with stochastic loss. *Journal of the Economic Science Association*, 1–15.

Charness, G., G. Genicot, et al. (2004). An experimental test of risk-sharing arrangements. *Economic Journal*.

Charness, G. and D. I. Levine (2007). Intention and stochastic outcomes: An experimental study. *The Economic Journal 117*(522), 1051–1072.

Chen, D. L., M. Schonger, and C. Wickens (2016). otreeâ€"an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance 9*, 88–97.

Chew, S. H., W. Wang, and S. Zhong (2023). Attention theory. *Working Paper*.

Coricelli, G., R. J. Dolan, and A. Sirigu (2007). Brain, emotion and decision making: the paradigmatic example of regret. *Trends in cognitive sciences 11*(6), 258–265.

Cosemans, M. and R. Frehen (2021). Salience theory and stock prices: Empirical evidence. *Journal of Financial Economics 140*(2), 460–483.

Costa-Gomes, M. A. and G. Weizsäcker (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies 75*(3), 729–762.

Danz, D. (2020). Never underestimate your opponent: Hindsight bias causes overplacement and overentry into competition. *Games and Economic Behavior 124*, 588–603.

Danz, D., D. Kübler, L. Mechtenberg, and J. Schmid (2015). On the failure of hindsight-biased principals to delegate optimally. *Management Science 61*(8), 1938–1958.

Danz, D., L. Vesterlund, and A. J. Wilson (2022). Belief elicitation and behavioral incentive compatibility. *American Economic Review 112*(9), 2851–2883.

Davis, B. J., R. Kerschbamer, and R. Oexl (2017). Is reciprocity really outcome-based? a second look at gift-exchange with random shocks. *Journal of the Economic Science Association 3*(2), 149–160.

de Oliveira, A. C., A. Smith, and J. Spraggon (2017). Reward the lucky? an experimental investigation of the impact of agency and luck on bonuses. *Journal of Economic Psychology 62*, 87–97.

Dertwinkel-Kalt, M., K. Köhler, M. R. Lange, and T. Wenzel (2017). Demand shifts due to salience effects: Experimental evidence. *Journal of the European Economic Association 15*(3), 626–653.

Dertwinkel-Kalt, M. and M. Köster (2015). Violations of first-order stochastic dominance as salience effects. *Journal of Behavioral and Experimental Economics 59*(2015), 42–46.

Dertwinkel-Kalt, M. and M. Köster (2017). Salient compromises in the newsvendor game. *Journal of Economic Behavior & Organization 141*, 301–315.

Dertwinkel-Kalt, M. and M. Köster (2019). Salience and skewness preferences. *Journal of the European Economic Association*.

Dertwinkel-Kalt, M. and M. Köster (2021). Replication: Salience and skewness preferences.

Diecidue, E. and J. Somasundaram (2017). Regret theory: A new foundation. *Journal of Economic Theory 172*, 88–119.

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.

Dufwenberg, M. and G. Kirchsteiger (2004). A theory of sequential reciprocity. *Games and economic behavior 47*(2), 268–298.

Eliaz, K. and R. Spiegler (2006). Contracting with diversely naive agents. *The Review of Economic Studies 73*(3), 689–714.

Engelbrecht-Wiggans, R. and E. Katok (2008). Regret and feedback information in first-price sealed-bid auctions. *Management Science 54*(4), 808–819.

Enke, B. and T. Graeber (2021). Cognitive uncertainty.

Enke, B., T. Graeber, and R. Oprea (2023). Confidence, self-selection, and bias in the aggregate. *American Economic Review 113*(7), 1933–1966.

Falk, A. and U. Fischbacher (2006). A theory of reciprocity. *Games and economic behavior 54*(2), 293–315.

Fehr, E., A. Klein, and K. M. Schmidt (2007). Fairness and contract design. *Econometrica 75*(1), 121–154.

Fehr, E. and K. M. Schmidt (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics 114*(3), 817–868.

Fehr, E. and J.-R. Tyran (2005). Individual irrationality and aggregate outcomes. *Journal of Economic Perspectives 19*(4), 43–66.

Filiz-Ozbay, E. and E. Y. Ozbay (2007). Auctions with anticipated regret: Theory and experiment. *American Economic Review 97*(4), 1407–1418.

Filiz-Ozbay, E. and E. Y. Ozbay (2014). Effect of an audience in public goods provision. *Experimental Economics 17*(2), 200–214.

Fisman, R., S. Kariv, and D. Markovits (2007). Individual preferences for giving. *American Economic Review 97*(5), 1858–1876.

Frazzini, A. and L. H. Pedersen (2014). Betting against beta. *Journal of financial economics 111*(1), 1–25.

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

Friedrichsen, J., K. Momsen, and S. Piasenti (2022). Ignorance, intention and stochastic outcomesâ˜†. *Journal of Behavioral and Experimental Economics 100*, 101913.

Frydman, C. and M. M. Mormann (2018). The role of salience in choice under risk: An experimental investigation.

Gabaix, X. (2014). A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics 129*(4), 1661–1710.

Gago, A. (2021). Reciprocity and uncertainty: When do people forgive? *Journal of Economic Psychology 84*, 102362.

Gauriot, R. and L. Page (2019). Fooled by performance randomness: Overrewarding luck. *Review of Economics and Statistics 101*(4), 658–666.

Gigerenzer, G. and U. Hoffrage (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological review 102*(4), 684.

Gollier, C. (2020). Aversion to risk of regret and preference for positively skewed risks. *Economic Theory 70*(4), 913–941.

Gollier, C. and B. Salanié (2006). Individual decisions under risk, risk sharing and asset prices with regret. *Working Paper*.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association 1*(1), 114–125.

Grossmann, M. (2023). Otree slider. Technical report.

Guiso, L. (2015). A test of narrow framing and its origin. *Italian Economic Journal 1*(1), 61–100.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica: Journal of the Econometric Society*, 667–686.

Gurdal, M. Y., J. B. Miller, and A. Rustichini (2013). Why blame? *Journal of Political Economy 121*(6), 1205–1247.

Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in cognitive psychology 12*(3), 145.

Hart, S. (2005). Adaptive heuristics. *Econometrica 73*(5), 1401–1430.

Hart, S. and A. Mas-Colell (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68*(5), 1127–1150.

Healy, A. and N. Malhotra (2009). Myopic voters and natural disaster policy. *American Political Science Review 103*(3), 387–406.

Healy, A. and N. Malhotra (2013). Retrospective voting reconsidered. *Annual Review of Political Science 16*, 285–306.

Healy, A. J., N. Malhotra, and C. H. Mo (2010). Irrelevant events affect voters' evaluations of government performance. *Proceedings of the National Academy of Sciences 107*(29), 12804–12809.

Hertwig, R., G. Barron, E. U. Weber, and I. Erev (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science 15*(8), 534–539.

Herweg, F. and D. Müller (2021). A comparison of regret theory and salience theory for decisions under risk. *Journal of Economic Theory 193*, 105226.

Heuer, J., C. Merkle, and M. Weber (2017). Fooled by randomness: Investor perception of fund manager skill. *Review of Finance 21*(2), 605–635.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, 74–91.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American economic review 92*(5), 1644–1655.

Hossain, T. and R. Okui (2013). The binarized scoring rule. *Review of Economic Studies 80*(3), 984–1001.

Hsee, C. K. and E. U. Weber (1999). Cross-national differences in risk preference and lay predictions. *Journal of Behavioral Decision Making 12*(2), 165–179.

Humphrey, S. J. (1995). Regret aversion or event-splitting effects? more evidence under risk and uncertainty. *Journal of Risk and Uncertainty 11*(3), 263–274.

Jenter, D. and F. Kanaan (2015). Ceo turnover and relative performance evaluation. *the Journal of Finance 70*(5), 2155–2184.

Kahneman, D. and A. Tversky (1979). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–292.

Karceski, J. (2002). Returns-chasing behavior, mutual funds, and beta's death. *Journal of Financial and Quantitative analysis 37*(4), 559–594.

Kessler, J. and L. Vesterlund (2015). The external validity of laboratory experiments: The misleading emphasis on quantitative effects. *Handbook of experimental economic methodology 18*, 392–405.

Kling, L., C. König-Kersting, and S. T. Trautmann (2023). Investment preferences and risk perception: Financial agents versus clients. *Journal of Banking & Finance 154*, 106489.

Kneer, M. and E. Machery (2019). No luck for moral luck. *Cognition 182*, 331–348.

Koller, T., D. Lovallo, and Z. Williams (2012). Overcoming a bias against risk. *McKinsey Quarterly 4*, 15–17.

König-Kersting, C., M. Pollmann, J. Potters, and S. T. Trautmann (2021). Good decision vs. good results: Outcome bias in the evaluation of financial agents. *Theory and Decision 90*(1), 31–61.

Königsheim, C., M. Lukas, and M. Nöth (2019). Salience theory: Calibration and heterogeneity in probability distortion. *Journal of Economic Behavior & Organization 157*, 477–495.

Kőszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118.

Krieger, J., D. Li, and D. Papanikolaou (2022). Missing novelty in drug development. *The Review of Financial Studies 35*(2), 636–679.

Lanzani, G. (2022). Correlation made simple: Applications to salience and regret theory. *The Quarterly Journal of Economics 137*(2), 959–987.

Lefgren, L., B. Platt, and J. Price (2015). Sticking with what (barely) worked: A test of outcome bias. *Management Science 61*(5), 1121–1136.

Leland, J. W. (1998). Similarity judgments in choice under uncertainty: A reinterpretation of the predictions of regret theory. *Management Science 44*(5), 659–672.

Leland, J. W., M. Schneider, and N. T. Wilcox (2019). Minimal frames and transparent frames for risk, time, and uncertainty. *Management Science 65*(9), 4318–4335.

Levitt, S. D. and J. A. List (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic perspectives 21*(2), 153–174.

Loewenfeld, M. (2024). Outcome bias and risk taking in a principal agent setting. *Working Paper*.

Loewenfeld, M. and J. Zheng (2023). Uncovering correlation sensitivity in decision making under risk. Technical report, Working Paper.

Loewenfeld, M. and J. Zheng (2024). Salience or event-splitting? an experimental investigation of correlation sensitivity in risk-taking. *Journal of the Economic Science Association*, 1–21.

Loomes, G. (1988a). Further evidence of the impact of regret and disappointment in choice under uncertainty. *Economica 55*(217), 47–62.

Loomes, G. (1988b). When actions speak louder than prospects. *American Economic Review*, 463–470.

Loomes, G. (2005). Modelling the stochastic component of behaviour in experiments: Some issues for the interpretation of data. *Experimental Economics 8*, 301–323.

Loomes, G., C. Starmer, and R. Sugden (1989). Preference reversal: information-processing effect or rational non-transitive choice? *The Economic Journal 99*(395), 140–151.

Loomes, G., C. Starmer, and R. Sugden (1991). Observing violations of transitivity by experimental methods. *Econometrica*, 425–439.

Loomes, G. and R. Sugden (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal 92*(368), 805–824.

Loomes, G. and R. Sugden (1987). Some implications of a more general form of regret theory. *Journal of Economic Theory 41*(2), 270–287.

Loomes, G. and R. Sugden (1995). Incorporating a stochastic element into decision theories. *European Economic Review 39*(3-4), 641–648.

Lovallo, D., T. Koller, R. Uhlaner, and D. Kahneman (2020). Your company is too risk-averse. *Harvard Business Review 98*(2), 104–124.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology 46*(1), 1–27.

Madarász, K. (2012). Information projection: Model and applications. *The Review of Economic Studies 79*(3), 961–985.

Mao, J. C. (1970). Survey of capital budgeting: Theory and practice. *Journal of Finance 25*(2), 349–360.

Martin, J. W. and F. Cushman (2016). Why we forgive what canâ€™t be controlled. *Cognition 147*, 133–143.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. *Frontier in Econometrics*, 105–142.

McLachlan, G. J., S. X. Lee, and S. I. Rathnayake (2019). Finite mixture models. *Annual review of statistics and its application 6*, 355–378.

Michenaud, S. and B. Solnik (2008). Applying regret theory to investment choices: Currency hedging decisions. *Journal of International Money and Finance 27*(5), 677–694.

Neumann, J. v. and O. Morgenstern (1953). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Ostermair, C. (2021). Investigating the empirical validity of salience theory: The role of display format effects. *Available at SSRN 3903649*.

Ostermair, C. (2022). An experimental investigation of the allais paradox with subjective probabilities and correlated outcomes. *Journal of Economic Psychology 93*, 102553.

Politis, D. N. and J. P. Romano (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 2031–2050.

Pollmann, M. M., J. Potters, and S. T. Trautmann (2014). Risk taking by agents: The role of ex-ante and ex-post accountability. *Economics Letters 123*(3), 387–390.

Polman, E. and K. Wu (2020). Decision making for others involving risk: A review and meta-analysis. *Journal of Economic Psychology 77*, 102184.

Pronin, E. (2007). Perception and misperception of bias in human judgment. *Trends in cognitive sciences 11*(1), 37–43.

Pronin, E., D. Y. Lin, and L. Ross (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin 28*(3), 369–381.

Qin, J. (2015). A model of regret, investor behavior, and market turbulence. *Journal of Economic Theory 160*, 150–174.

Quiggin, J. (1990). Stochastic dominance in regret theory. *The Review of Economic Studies 57*(3), 503–511.

Quiggin, J. (1994). Regret theory with general choice sets. *Journal of Risk and Uncertainty 8*(2), 153–165.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American economic review*, 1281–1302.

Regenwetter, M., J. Dana, and C. P. Davis-Stober (2011). Transitivity of preferences. *Psychological Review 118*(1), 42.

Romano, J. P. and M. Wolf (2005). Stepwise multiple testing as formalized data snooping. *Econometrica 73*(4), 1237–1282.

Romano, J. P. and M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters 113*, 38–40.

Rubin, J. and R. Sheremeta (2016). Principal–agent settings with random shocks. *Management Science 62*(4), 985–999.

Russell, T. and R. Thaler (1985). The relevance of quasi rationality in competitive markets. *The American Economic Review 75*(5), 1071–1082.

Samek, A. and J. R. Sydnor (2020). Impact of consequence information on insurance choice. Technical report, National Bureau of Economic Research.

Sarver, T. (2008). Anticipating regret: Why fewer options may be better. *Econometrica 76*(2), 263–305.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

Selten, R. (1967). Die strategiemethode zur erforschung des eingeschr nkt rationale verhaltens im rahmen eines oligopolexperiments. *Beitr ge zur experimentellen Wirtschaftsforschung*, 136.

Sezer, O., T. Zhang, F. Gino, and M. H. Bazerman (2016). Overcoming the outcome bias: Making intentions matter. *Organizational Behavior and Human Decision Processes 137*, 13–26.

Somasundaram, J. and E. Diecidue (2017). Regret theory and risk attitudes. *Journal of Risk and Uncertainty 55*(2-3), 147–175.

Sopher, B. and G. Gigliotti (1993). Intransitive cycles: Rational choice or random error? an answer based on estimation of error rates with experimental data. *Theory and Decision 35*, 311–336.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature 38*(2), 332–382.

Starmer, C. and R. Sugden (1993). Testing for juxtaposition and event-splitting effects. *Journal of Risk and Uncertainty 6*(3), 235–254.

Strack, P. and P. Viefers (2021). Too proud to stop: Regret in dynamic decisions. *Journal of the European Economic Association 19*(1), 165–199.

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

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty 5*(4), 297–323.

Van Leeuwen, B. and I. Alger (2019). Estimating social preferences and kantian morality in strategic interactions.

Wakker, P. and D. Deneffe (1996). Eliciting von neumann-morgenstern utilities when probabilities are distorted or unknown. *Management Science 42*(8), 1131–1150.

Wong, K. P. (2012). Production and insurance under regret aversion. *Economic Modeling 29*(4), 1154–1160.

Zeelenberg, M. (1999). Anticipated regret, expected feedback and behavioral decision making. *Journal of Behavioral Decision Making 12*(2), 93–106.

Zeelenberg, M., J. Beattie, J. Van der Pligt, and N. K. De Vries (1996). Consequences of regret aversion: Effects of expected feedback on risky decision making. *Organizational Behavior and Human Decision Processes 65*(2), 148–158.

Zheng, J. (2021). Willingness to pay for reductions in health risks under anticipated regret. *Journal of Health Economics 78*(2021), 102476.

Appendix A

Appendix to Chapter 1

Appendix A: Proof of Proposition 1

Step 1 and 2 of the proof follow that of Lanzani (2022) closely. However, we replace the condition of modularity of the ϕ function with the CSPD property.

Step 1: (3) $\implies$ (2). Take any $h, m, l \in X$. Without loss of generality, suppose $h \geq m \geq l$, and define $u : X \to \mathbb{R}$ such that $u(z) = \phi(z, l)$. According to the statement (3) and by definition of $u(\cdot)$, we have

$$\phi(h,l) = \phi(h,m) + \phi(m,l) \iff \phi(h,m) = \phi(h,l) - \phi(m,l) = u(h) - u(m).$$

This proves that $\phi(h,m) = u(h) - u(m)$ whenever $h \geq m$. Whenever $m > h$, skew-symmetry of $\phi(\cdot)$ implies $\phi(h,m) = -\phi(m,h) = -(u(m) - u(h)) = u(h) - u(m)$. This proves that $\phi(h,m) = u(h) - u(m)$ holds for any $m, h \in X$.

Thus, $\pi \in \Pi$ if and only if

$$\sum_{(x,y)\in X\times X} \phi(x,y)\pi(x,y) \geq 0 \iff \sum_{(x,y)\in X\times X} [u(x) - u(y)]\pi(x,y) \geq 0 \iff$$
$$\sum_{x\in X} \pi_1(x)u(x) \geq \sum_{x\in X} \pi_2(x)u(x).$$

Step 2: (2) $\implies$ (3). If Π admits an expected utility representation, then

$$\pi \in \Pi \iff \sum_{(x,y)\in X\times X} (u(x) - u(y))\pi(x,y).$$

Defining $\phi(z,w) = u(z) - u(w)$, (3) is satisfied for any $h, m, l \in X$, that is

$$\phi(h,l) = \phi(h,m) + \phi(m,l) \iff u(h) - u(l) = u(h) - u(m) + u(m) - u(l).$$

Step 3: (2) $\implies$ (1) is a well known result.

To proof (1) $\implies$ (3), we first proof the following lemma, which clarifies that a strict preference for one of the lotteries of the SML task constitutes a violation of transitivity in a quite general framework where only the completeness axiom is imposed.

Lemma A. When the completeness axiom is satisfied, (1) π satisfies transitivity $\implies$ (a) $\pi, \overline{\pi} \in \Pi$ for any $\pi \in \Delta(X \times X)$ such that $\pi_1 = \pi_2$.

We prove (1) $\implies$ (a) by contrapositive, that is not (a) $\implies$ not (1). Not (a) implies that there exists π with $\pi_1 = \pi_2$ such that $\pi \in \hat{\Pi}$. Consider also the joint distribution χ with $\chi_1 = \chi_2 = \pi_1$ such that $\chi = \overline{\chi}$. Completeness implies $\chi, \overline{\chi} \in \Pi$. Finally, define the joint distribution $\rho = \overline{\pi}$. Preferences over the three joint distributions π, χ, and ρ violate transitivity.

For the reader's convenience, we restate here the transitivity axiom. Transitivity means that $\forall\ \pi, \chi, \rho \in \Delta(X \times X)$, if $\pi_2 = \chi_1$, $\rho_1 = \pi_1$, and $\rho_2 = \chi_2$, then $(\pi \in \Pi, \chi \in \Pi) \Rightarrow \rho \in \Pi$. In the example, $\pi_1 = \pi_2 = \rho_1 = \rho_2 = \chi_1 = \chi_2$, and $(\pi \in \Pi, \chi \in \Pi)$. But, by construction, $\rho \notin \Pi$. This concludes the proof of the lemma.

Step 4: (1) $\implies$ (3). We prove this by contrapositive, that is showing that not (3) $\implies$ not (1). Consider $\pi = ((h,l), 1/3; (m,h), 1/3; (l,m), 1/3)$, where $h > m > l$. Note that $\pi_1 = \pi_2$. Imposing $\phi(h,l) > \phi(h,m) + \phi(m,l)$ or $\phi(h,l) < \phi(h,m) + \phi(m,l)$ implies $\pi \in \hat{\Pi}$ or $\overline{\pi} \in \hat{\Pi}$ respectively. By Lemma A.1, this violates transitivity.

This concludes the proof proposition 1.

It is worth pointing out that the only way to violate transitivity in Lanzani's framework is to have preferences such that $\phi(h,l) \neq \phi(h,m) + \phi(m,l)$. Lemma A.1 shows that such preferences constitute a violation of transitivity in a much more general framework that imposes only the completeness axiom.

Appendix B

Appendix to Chapter 3

Appendix A: Related Literature

First and foremost, my paper contributes to a growing experimental literature that studies how OB shapes ex-post rewards and punishments in principal-agent settings. The study of OB has been sparked by the seminal paper of Baron and Hershey (1988), who document OB in decision evaluation in a series of vignette studies.[1] In an important paper, Gurdal et al. (2013) demonstrate violations of the informativeness criteria in setting of delegated risk-taking. They find that ex-post reward decisions do not only depend on agents' choices but also on the obtained as well as the forgone outcome, even when principals can observe choices perfectly. These outcome effects are robust to learning and also occur for bonus decisions made by independent third parties. de Oliveira et al. (2017) consider OB in reward decisions in a setting of investment decisions. They find that whereas luck is not rewarded with higher bonuses, bad luck induces lower bonus payments, compared to a baseline analogous to the reward-before condition in this paper. König-Kersting et al. (2021) consider a similar setting. They find OB in reward decisions as well as in ratings of decision quality. Moreover, OB maintains a strong influence, even when the agent is informed about the principal's preferred investment level. Aimone and Pan (2020) demonstrate the OB in a setting of delegated risk-taking that is augmented with a trust-game stage in which the principal decides how much of her endowment to give to the agent for investment. Brownback and Kuhn (2019) study OB in a setting of costly effort provision. They demonstrate that OB can crowd out the principals' tendency to reward based on the agent's effort. They also show that agents can anticipate the principals' OB and are willing to pay in order to manipulate the information available to the principal prior to their making a bonus decision.

I advance this literature by studying the implications of OB rather than focusing on documenting it and studying its causes. I build on the finding of Gurdal et al. (2013) that ex-post reward decisions are strongly impacted by an ex-post counterfactual comparison and combine it with ideas from the decision-theoretic literature on regret theory (Loomes and Sugden, 1982, 1987; Quiggin,

[1] A subsequent literature in psychology provides consistent evidence that OB can impact ex-post decision evaluation (Boles and Messick, 1995) and moral judgment (Sezer et al., 2016; Kneer and Machery, 2019; Martin and Cushman, 2016).

1990). This enables me to provide a simple yet powerful framework to think about how the OB impacts agents' implicit incentives to choose between different risks and, ultimately, the agent's choice and welfare.

A second strand of experimental papers studies whether OB can disrupt welfare improving reciprocal relationships between agent and principal. Charness and Levine (2007) do not find much evidence that principals exhibit OB or that it can indeed distort the reciprocal relationship between agent and principal. Rubin and Sheremeta (2016) consider a setting in which agents have a convex cost function. They find that OB makes bonus decisions more random, which induces agents to lower their effort in response. However, Davis et al. (2017) fail to replicate these results and argue that the findings of Rubin and Sheremeta (2016) might be attributable to incidental differences in behavior ("caused by subject pool effects or by pure chance"(Davis et al., 2017, 158)) between treatments early in the experiment that lead to a hysteresis effect.

I contribute to this literature by providing a new lens through which to study the implications of OB. Importantly, my model provides a natural explanation for why the observed effects of OB do not seem very strong and stable in the considered settings. In all settings, one of the agent's actions dominates the other state-wise, that is, it always yields a (weakly) better outcome than the alternative. In my model, this implies that the ex-post counterfactual comparison is always aligned with the ex-ante comparison, which limits the scope of OB to cause harm. This highlights the usefulness of ex-post counterfactual to understand when OB is likely to matter and when it is not. Another crucial difference to the present study is that I consider a setting without any of the classical agency problems and show that reciprocity, when based on the ex-post counterfactual comparison, can create a new kind of agency problem rather than alleviating existing moral hazard problems.

My paper is further closely linked to the experimental literature on delegated risk-taking. Much research on delegated risk-taking focuses on the question of whether agents take more or less risk for others than for themselves. Polman and Wu (2020) conduct a meta-analysis and find that overall, agents take somewhat more risks for others than for themselves, though there is a considerable number of studies finding the opposite tendency. The authors conclude that research should focus on determining when delegated risk-taking might induce a risky and when it might induce a safe shift. Among the literature on delegated risk-taking, the study closest to my paper might be Pollmann et al. (2014) which study how agents' level of risky investment in a repeated investment game differs depending on whether principals decide on their bonus payments in conditions analogous to the reward-before and reward-after treatment discussed above. They find that overall investment levels in their reward-after condition are comparable to investment levels chosen by participants for themselves, whereas investment levels in the reward-before condition are much lower.

My paper contributes to this literature by showing how ex-post counterfactual comparison might systematically impact the risks chosen by agents. My findings suggest that ex-post counterfactual evaluation could lead to either more or less risk-taking, depending on which risks are more likely to yield a higher outcome ex-post.

Another related strand of literature is on behavioral aspects of principal-agent interactions.

Kőszegi (2014) reviews the literature on behavioral contract theory. Much of this literature focuses on how rational principals change the optimal contract when dealing with "behavioral" agents, for instance, agents that have fairness concerns (Fehr et al., 2007) or are present biased (Eliaz and Spiegler, 2006). Closer to this paper, a growing literature investigates the implications of biases arising at the principal's side. Danz et al. (2015) provides experimental evidence that hindsight-biased principals might underestimate the skill of agents and, as a result, delegate too little. Madarász (2012) provides a theoretical treatment of information projection, a tendency to project one's own information set onto that of others. He considers applications to problems of moral hazard and shows theoretically that information projection at the time of evaluation onto the agent's information set at the time of decision-making can create agency problems.[2] I complement these studies by showing how ex-post counterfactual comparison of outcomes can induce agency problems when the agent chooses between certain types of risks.

Appendix B: Model

B.2.1 Injecting outcome bias into a reciprocity model

I consider the workhorse model of reciprocity developed by Rabin (1993) for normal form games and extended to sequential games by Dufwenberg and Kirchsteiger (2004) (the RDK model henceforth). There are two important terms capturing the second mover's reciprocity motive. The kindness term $K(bonus)$ captures how kind the second mover is to the first mover and is a function of the principal's bonus decision. The perceived kindness term $\varphi(\theta, s)$ captures how kind the second mover perceives the action of the first mover to be. To consider outcome effects, I allow the principal's kindness perception to depend both on the agent's lottery choice θ and the realized state s. I discuss shortly how these terms are defined. The principal's utility function is given by

$$U_p = v(x_s^\theta) + \rho K(bonus)\varphi(\theta, s), \tag{B.1}$$

where the parameter $\rho > 0$ captures the strength of the principal's reciprocity motive. If $\varphi(\theta, s) > 0$ (< 0), the principal perceives the agent to be kind (unkind), and her utility is increasing (decreasing) in $K(bonus)$. This captures the principal's reciprocity motive.[3]

The principal's kindness towards the agent is defined as $K(bonus) = M_a(bonus) - M_a^E(bonus)$ where $M_a(bonus)$ is the payoff the agent receives, and $M_a^E(bonus)$ is a reference point relative to which the agent's final payoff is evaluated. Dufwenberg and Kirchsteiger (2004) define this reference point to be the midpoint between the highest and lowest payoff the second mover can give to the first

[2] OB could be interpreted as an extreme case of information projection. In this interpretation, the principal, at the time of evaluating the agent's choice, puts some weight on the agent knowing the outcomes of the lotteries at the time of decision-making. While this interpretation is possible, it seems somewhat counter-intuitive.

[3] Note that the passive player has been omitted from the principal's utility function. As the passive player cannot impact the principal's payoff, they are perceived as neither kind nor unkind, that is, $\varphi(\theta, s) = 0$.

mover by choosing an action among a set of efficient actions.[4] In the current setting, this implies that the reference point is defined as $M_a^E(bonus) = 0.5[max\{M_a(bonus)\} + min\{M_a(bonus)\}] = 0.5$. Therefore $K(bonus) = bonus - 0.5$. In words, the principal is kind to the agent if they award the bonus and not otherwise.

Perceived kindness is defined as the payoff the first mover intends to give to the second mover, relative to a reference point, which is given by the midpoint between the highest and lowest payoff the first mover can give to the second mover.[5] In the absence of OB, the perceived kindness term, $\varphi(\theta)$, is a function of the agent's lottery choice only. Dufwenberg and Kirchsteiger (2004) restrict attention to games without nature. In the present setting, nature determines which state of the world materializes, which requires imposing assumptions about what determines the principal's kindness perception. It seems natural to define perceived kindness based on the expected utility the first mover intended to give to the second mover, relative to the reference point.

Denote $EU^\theta = \sum_s q_s u(x_s^\theta)$ the expected utility of lottery θ. Then, $\varphi(\theta, s) = EU^\theta - EU^E$, where $EU^E = 0.5[max_\theta\{EU^\theta\} + min_\theta\{EU^\theta\}]$. Since the agent can choose between only two lotteries, $\varphi(\theta, s) = 0.5[EU^\theta - EU^{-\theta}]$, where $-\theta$ denotes the non-chosen lottery. The principal thus considers the agent as kind if he chose her preferred lottery and unkind otherwise, which again seems very natural.[6]

It seems natural to define the outcome-biased perceived kindness by replacing the expected utility in the expression derived above by its outcome-biased equivalent. Then, $\varphi(\theta, s) = OBEU_s^\theta - OBEU_s^E$, where $OBEU_s^E = 0.5[max_\theta\{OBEU_s^\theta\} + min_\theta\{OBEU_s^\theta\}]$. Since the agent can choose between only two lotteries,

$$\varphi(\theta, s) = 0.5[\lambda_p \underbrace{[u(x_s^\theta) - u(x_s^{-\theta})]}_{\text{Ex-post comparison of outcomes}} + (1 - \lambda_p) \underbrace{[EU^\theta - EU^{-\theta}]}_{\text{Ex-ante comparison of EU}}] \tag{B.2}$$

That is, the principal's kindness perception is simply the sum of an ex-ante comparison of the difference in expected utilities and an ex-post comparison of the realized utilities of the chosen and non-chosen lottery, weighted by the principal's degree of OB λ_p. Note that, for $\lambda_p = 0$, $\varphi(\theta, s) = \varphi(\theta)$, that is, the perceived kindness is a function of the agent's lottery choice only.

It follows from the principal's utility function that her bonus decisions adhere to a simple rule: award the bonus to the agent whenever $\varphi(\theta, s) \geq 0$ and not otherwise. This yields expression 4.3.

[4] Actions are restricted to a set of efficient actions such that, for any history of play, there is no other action that provides a strictly higher payoff to at least one player and no lower payoff for other players. Since the principal chooses between giving the bonus to the agent and giving it to a passive third party, both $bonus = 0$ and $bonus = 1$ are in the set of efficient actions.

[5] Dufwenberg and Kirchsteiger (2004) consider settings in which the second mover's payoff depends on her own choice. The payoff the first mover intends to give to the second mover is thus a function of the first mover's beliefs about the actions of the second mover, and the perceived kindness is a function of the second mover's beliefs over these beliefs (i.e., the second mover's second-order beliefs). Since the principal cannot impact her own payoff in the present setting, no second-order beliefs have to be considered. Since the principal's bonus decision terminates the interaction, her first-order beliefs are irrelevant. This simplifies the analysis considerably.

[6] Complications might arise when it is not clear which lottery the principal prefers. In the experiment, this issue will be circumvented by having agents choose between pairs of first-order-stochastic dominant and dominated lotteries.

B.2.2 Proofs of proposition 3

I reproduce the proposition below for the convenience of the reader.

Proposition 4. *Suppose* $EU^G > EU^B$.

a) $I(\lambda_p)$ *is weakly decreasing in* λ_p*;*

b) Whenever $\sum_{s=1}^{S} q_s \mathbb{1}\{\Delta_s < 0\} > 0.5$, $I(\lambda_p) < 0$ *for all* $\lambda_p \in (\underline{\lambda_p}, 1]$*;*

c) For any lottery G *with at least 3 distinct outcomes and* $\max_s\{q_s\} < 0.5$*, there exists a lottery* B*, such that* G *is first-order-stochastic dominant to* B*, but* $I(\lambda_p) < 0$ *for all* $\lambda_p \in (\underline{\lambda_p}, 1]$.

Statement a): Whenever $\Delta_s > 0$, the principal awards the bonus for lottery G, but not B, for all $\lambda_p \in [0, 1]$. Whenever $\Delta_s < 0$, the principal awards the bonus for lottery G if her degree of OB λ_p is below a threshold level, that is if $\lambda_p < \frac{\Delta(EU)}{\Delta(EU)-\Delta_s}$. Thus, the incentives to choose lottery G can be written as $I(\lambda_p) = 2\sum_{s=1}^{S} q_s \mathbb{1}\{\lambda_p < \frac{\Delta(EU)}{\Delta(EU)-\Delta_s}\} - 1$. Whenever $\Delta_s < 0$ for at least one state, $I(\lambda_p)$ is thus weakly decreasing in $\lambda_p \in [0, 1]$.

Statement b): $I(\lambda_p) < 0$ whenever $Pr(\lambda_p) = \sum_{s=1}^{S} q_s \mathbb{1}\{\Delta_s < 0\} < 0.5$. Order the S states of the world such that $\Delta_1 \leq \Delta_2 \leq ... \leq \Delta_S$. Define the state k such that $\Delta_k \in \{\max_s : \sum_{s=1}^{k} q_s > 0.5\}$. Then, $\underline{\lambda_p} = \frac{\Delta(EU)}{\Delta(EU)-\Delta_k}$, defines the threshold values such that $Pr(\lambda_p) < 0.5$ for all $\lambda \in (\underline{\lambda_p}, 1]$

Statement c): For the proof of statement c), note that, given statement b), it suffices to show that for every lottery G satisfying the stated conditions, one can find a lottery B that is 1) first-order-stochastically dominated by G, and 2) $\sum_{s\in S} q_s \mathbb{1}\{\Delta_s < 0\} > 0.5$. In the following, I show that for a lottery G with $k \in K$ distinct outcomes, $max \sum_{s\in S} q_s \mathbb{1}\{\Delta_s < 0\} = 1 - max_k\{q_k\}$, which yields the result immediately. The proof is inspired by (Quiggin, 1994) and is instructive in that it highlights the fundamental reason for the agency problem that arises from the principal's ex-post counterfactual comparison.

Consider a lottery G with K distinct payoffs y_k that each occurs with probability q_k. Suppose payoffs are ordered such that $y_k > y_{k+1}$. Using a trick from Quiggin (1990), this marginal distribution can be represented by a lottery with S equiprobable states, such that each payoff occurs in N_k states such that $q_k = \frac{N_k}{S}$.[7] Now, arrange the states in order of the payoff of lottery G, that is $x_1^G > x_2^G > ... > x_S^G$. Next, define a lottery B such that $x_s^B = x_s^G - \epsilon$, where $0 < \epsilon < min_k y_k - y_{1+k}$. Consider a lottery B'. Following Quiggin (1990), define $\nu(\cdot)$ to be a bijection of $\{1, ..., S$ onto itself. Lottery B' is obtained from lottery B by permuting payoffs across state, that is $x_{\nu(s)}^{B'} = x_s^B$. The permutation function $\nu(\cdot)$ satisfies $\nu(\nu(s)) = s$. Obviously, lottery G is first-order-stochastic-dominant to lottery B' for any such permutation.

Consider the following algorithm that describes a permutation $\nu^*(\cdot)$ of payoffs. Start at $s = 1$, then proceed to state $s = 2$, etc. Consider state $j \in \{1, ..., S\}$. Define $\hat{s} \in \hat{S} = s \notin \{\nu^*(1), ..., \nu^*(j-1)\}$, i.e. the set of states that no payoff has been permuted to thus far. Then, for state $j \in S$

1. set $\nu^*(j) = min\{\hat{s} : x_j^B > x_{\hat{s}}^G\}$

[7] This requires the probabilities q_k to be rational numbers. If they are not, they can be approximated to an arbitrary degree.

2. if $\nexists$ such $\hat{s}$, set $\nu^*(j) = s \in \{\hat{s} : x_{\hat{s}}^G > x_j^B\}$

This algorithm ensures that $\nu^*(\cdot) = \underset{\nu(\cdot)}{argmax} \sum_{s=1}^{S} \mathbb{1}\{x_s^G < x_{\nu(s)}^B\}$ by starting with the highest payoff of lottery B, and permuting it to a state in which lottery G yields the highest possible payoff that is lower than the considered payoff of lottery B. Whenever this is not possible, the considered payoff of B is permuted into a state where lottery G yields a higher payoff than the considered payoff of lottery B.

I now show that $\sum_{s \in S} q_s \mathbb{1}\{x_s^G < x_{\nu^*(s)}^B\} = 1 - max\{q_k\}$. First denote $j_k = min\{s : x_s^G = y_k\}$, i.e. the lowest state in which lottery G yields a payoff y_k. Consider, $\nu^{*^{-1}}(j_k)$,the state whose payoff x_s^B is permuted to state state j_k for lottery B' . There are $j_k - \nu^{*^{-1}}(j_k)$ states in which $x_s^B > y_k$ that can be permuted to a state in which $x_s^G = y_k$. Therefore,

$$\sum_{s \in \{j_k, j_k+1, \ldots, j_k+N_k\}} \mathbb{1}\{x_s^G < x_{\nu^*(s)}^B\} = min\{j_k - \nu^{*^{-1}}(j_k), N_k\} \tag{B.3}$$

Next note that for $k > 1$,

$$j_k - \nu^{*^{-1}}(j_k) = max\{N_{k-1}, j_{k-1} - \nu^{*^{-1}}(j_{k-1})\} \tag{B.4}$$

It follows that for $k > 1$,

$$j_k - \nu^{*^{-1}}(j_k) = max_{k \in \{1, \ldots, j_k - 1\}} N_k \tag{B.5}$$

To see this, not that for , $j_2 - \nu^{*^{-1}}(j_2) = N_1$. Therefore, $j_3 - \nu^{*^{-1}}(j_3) = max\{N_2, N_1\ \}$, etc.

Combining equations B.3 and B.5 yields

$$\sum_{s \in S} \mathbb{1}\{x_s^G < x_{\nu^*(s)}^B\} = \sum_{k \in \{1, \ldots, K\}} min\{N_k, max_{i \in \{1, \ldots, k-1\}}\{N_i\}\} = 1 - max_{k \in \{1, \ldots, K\}} N_k, \tag{B.6}$$

or equivalently $\sum_{s \in S} q_s \mathbb{1}\{x_s^G < x_{\nu^*(s)}^G\} = 1 - max\{q_k\}$. $\square$

B.2.3 Link to correlation sensitivity

B.2.3.1 Representation as correlation-sensitive preference

In this section, I provide a formal argument that the fundamental reason why OB induces incentives to deviate from the principal's ex-ante preferences is that it induces incentives to be correlation-sensitivity. For the sake of the exercise, I distinguish between the principal's ex-ante preferences and her preferences as expressed by the incentives that arise due to outcome-biased bonus decisions. I show that these expressed preferences can be represented by a utility function that satisfies the axioms of the general theory of correlation sensitivity proposed by Lanzani (2022).

First, some pieces of notation need to be introduced. In Lanzani's framework, preferences are defined over the joint distribution of payoffs rather than binary relations over lotteries defined by

the help of introducing a state-space. Lottery choices are described by a nonempty set of payoffs X and a finite measure of the joint probability distribution $\pi \in \Delta(X \times X)$. To avoid technicalities, I consider $X \subseteq \mathbb{R}$. The decision maker has to choose a lottery $\theta \in \{G, B\}$ so as to be paid accordingly to the realized payoff of lottery θ. The payoff pair (g_i, b_j) realizes with a probability of π_{ij}. Lanzani defines the decision maker's preferences over the joint distribution of outcomes. Binary preferences are modeled as a preference set $\Pi \subseteq \Delta(X \times X)$. The decision maker is said to have a preference for lottery G for a given joint distribution π if $\pi \in \Pi$. Define $\overline{\pi}$ to be the conjugate distribution of π, that is, $\forall (g, b) \in X \times X$, $\overline{\pi}(g, b) = \pi(b, g)$. Intuitively, the conjugate distribution is the distribution that obtains when relabeling lottery G (calling it B) and B (calling it G). Then, a decision maker has a preference for lottery B if $\overline{\pi} \in \Pi$.

The relation $\pi \in \Pi$ is analogous to the familiar weak preference relation $\succeq$, and $\pi, \overline{\pi} \in \Pi$ corresponds to indifference.[8] Given the notion of weak preferences, a notion of strict preferences in the language of preference sets can be introduced. Given a preference set Π, the subset of strict preferences $\hat{\Pi}$ is defined as $\hat{\Pi} = \{\pi \in \Pi : \overline{\pi} \notin \Pi\}$. That is, for $\pi \in \hat{\Pi}$, the decision maker strictly prefers to be paid according to lottery G rather than lottery B.

Lanzani (2022) imposes three axioms on the preference set Π that are necessary and sufficient to obtain a representation theorem for correlation-sensitive preferences that I will return to shortly. Consider a skew-symmetric function $\phi : X \times X \to \mathbb{R}$, that is, $\phi(g, b) = -\phi(b, g)$, $\forall (g, b) \in \mathbb{R} \times \mathbb{R}$. A preference set Π satisfies Lanzani's three axioms if and only if there exists a skew-symmetric function ϕ such that, for any $\pi \in \Delta(X \times X)$

$$\pi \in \Pi \iff \sum_{g,b} \phi(g, b)\pi(b, g) \geq 0 \tag{B.7}$$

I can now consider the principal's expressed preferences. I say that a principal with a degree of OB λ expresses a preference through her ex-post bonus decisions for lottery G whenever she is more likely to award the bonus for lottery G than for lottery B.

$$\pi_\lambda \in \Pi_\lambda \iff Pr(G, \lambda) \geq Pr(G, \lambda), \tag{B.8}$$

where the subscript λ highlights that these are not the principal's ex-ante preferences but her preferences as expressed through her outcome-biased ex-post bonus decisions.

I am now in the position to state the first result.

Proposition 5. *There exists a skew-symmetric function $\phi_\lambda : X \times X \to \mathbb{R}$ that satisfies*

$$\pi_\lambda \in \Pi_\lambda \iff \sum_{g,b} \phi_\lambda(g, b)\pi(b, g) \geq 0 \tag{B.9}$$

Proof: To proof proposition 5, on needs to show that the relation $\pi_\lambda \in \Pi_\lambda \iff Pr(G, \lambda)\}$

[8] The classical preference relation $\succeq$ also induces a preference set. However, it is defined over $\Delta(X) \times \Delta(X)$ whereas Π is defined over $\Delta(X \times X)$.

satisfies Lazani's axiom's ordering, strong independence, and archimedean continuity. I relegate this exercise to section B.2.3.3, where I also state the axioms formally.

The proposition states that the principal's expressed preferences can be represented by a correlation-sensitive preference function. The subscript λ highlights again that the function $\phi_\lambda(\cdot)$ represents the principal's expressed preferences.

B.2.3.2 Ex-ante vs expressed preferences

We are now in a position to see that OB drives a wedge between the principal's ex-ante preferences and her expressed preferences because OB induces the expressed preferences to be correlation-sensitive, even if her ex-ante preferences are insensitivity to the correlation structure. I say that preferences are correlation-insensitive if they can be fully described by a binary ranking over marginal distributions. In this is case, they admit an expected utility theory representation, as established by proposition 1 of Lanzani (2022).Formally, a preference set admits an EUT representation if there exists $u : X \to \mathbb{R}$ such that $\pi \in \Pi \iff \sum_{g,b}(u(g) - u(b))\pi(g,b) \geq 0$. Equivalently, $\phi(g,b) = u(g) - u(b)$.

Proposition 6. *Consider a principal whose ex-ante preferences are correlation-insensitive, that is, they admit an expected utility theory representation. The following statements are equivalent.*

1) $\pi \in \Pi \iff \pi_\lambda \in \Pi_\lambda$

2) Π_λ *admits an expected utility theory representation.*

3) $\lambda = 0$

Proof of proposition 6:

2) $\iff$ 3)

The proof makes use of proposition 1 of Loewenfeld and Zheng (2023) that shows that Π_λ admits an expected utility theory representation if and only if $\phi_\lambda(h,l) = \phi_\lambda(h,m) + \phi_\lambda(m,l)$, $\forall h > m > l \in \mathbb{R}$. Note that this statement is equivalent to indifference between two lotteries A and B with the joint distribution of payoffs $(1/3,(h,l);1/3,(m,h);1/3,(l,m))$. Note next that $Pr(G,\lambda) \geq Pr(B,\lambda) \iff \sum_{g,b}\Pi(g,b)\mathbb{1}\{\lambda(u(g)-u(b)) + (1-\lambda)\Delta(EU)\} \geq 0$. From imposing that the principal's ex-ante preferences are correlation-insensitive, it follows that $\Delta(EU) = 0$. To proof 2) $\iff$ 3), it then sufficient to note that

$$\mathbb{1}\{\lambda(u(h)-u(l))\} = \mathbb{1}\{\lambda(u(h)-u(m))\} + \mathbb{1}\{\lambda(u(m)-u(l))\} \iff \lambda = 0$$

1) $\iff$ 3)

I show 1) $\implies$ 3) by contra-positive, that is, showing that not 3) implies not 1). Consider a choice between two lotteries G and B with the following joint distribution $1/3,(h+\epsilon,l);1/3(m+\epsilon,h);1/3(l+\epsilon,m)$, where $\epsilon > 0$. Note that lottery G is first-order-stochastically dominant to lottery B and is therefore strictly preferred by any decision maker whose preferences admit an expected

utility theory representation, that is $\pi \in \hat{\Pi}$. Not 3) implies $\lambda > 0$. For any $\lambda > 0$, one can find an ϵ small enough such that the principal is strictly more likely to award the for the dominated lottery, that is $\pi_\lambda \notin \Pi$. This establishes 1) $\implies$ 3).

3) $\implies$ 1) is a trivial implication of the definition of expressed preferences. Recall that $Pr(G, \lambda) \geq Pr(B, \lambda) \iff \sum_{g,b} \Pi(g, b) \mathbb{1}\{\lambda(u(g) - u(b)) + (1 - \lambda)\Delta(EU)\} \geq 0$. For $\lambda = 0$, the principal awards the bonus if and only if $\Delta(EU) > 0$, which yields the result immediately.

This concludes the proof. □

B.2.3.3 Proof of proposition 5

I first state the axioms of Lanzani (2022).

Axiom 1 (Completeness): *For all* $\pi \in \Delta(X \times X)$

$$\pi \notin \Pi \implies \overline{\pi} \in \Pi$$

Axiom 2 (Strong independence): *For all* $\pi, \pi' \in \Pi$*, and all* $\alpha \in (0, 1)$

$$\alpha\pi + (1 - \alpha)\pi' \in \Pi$$

Moreover, if $\pi' \in \hat{\Pi}$*, then*

$$\alpha\pi + (1 - \alpha)\pi' \in \hat{\Pi}$$

Axiom 3 (Archimedean Continuity): *For all* $\pi \in \hat{\Pi}$*,* $\pi' \notin \Pi$*, there exists* $\alpha, \beta \in (0, 1)$ *such that*

$$\alpha\pi + (1 - \alpha)\pi' \in \hat{\Pi} \text{ and } \beta\pi + (1 - \beta)\pi' \notin \hat{\Pi}$$

To prove that the relation given in expression B.8 satisfies completeness, it is sufficient to note that $Pr(\theta, \lambda) \in [0, 1]$. This implies that, either $Pr(G, \lambda) \geq Pr(B, \lambda)$ or $Pr(G, \lambda) < Pr(B, \lambda)$.

To see that strong independence holds, note that $\pi, \pi' \in \Pi$ implies $Pr(G, \lambda) - Pr(B, \lambda) \geq 0$ and $Pr(G', \lambda) - Pr(B', \lambda) \geq 0$. It follows that $\alpha[Pr(G, \lambda) - Pr(B, \lambda)] + (1-\alpha)[Pr(G', \lambda) - Pr(B', \lambda)] \geq 0$ for all $\alpha \in (0, 1)$. Further, if $\pi' \in \hat{\Pi}$, then $Pr(G', \lambda) - Pr(B', \lambda) > 0$, which implies that $\alpha[Pr(G, \lambda) - Pr(B, \lambda)] + (1 - \alpha)[Pr(G', \lambda) - Pr(B', \lambda)] > 0$ for all $\alpha \in (0, 1)$.

Finally, to see that Archimedian Continuity holds, note that $\pi \in \hat{\Pi}$ implies $Pr(G, \lambda) - Pr(B, \lambda) > 0$ and $\pi' \notin \Pi$ implies $Pr(G', \lambda) - Pr(B', \lambda) < 0$. Then , $\alpha[Pr(G, \lambda) - Pr(B, \lambda)] + (1-\alpha)[Pr(G', \lambda) - Pr(B', \lambda)] \geq 0$ for $\alpha \geq \frac{-[Pr(G', \lambda) - Pr(B', \lambda)]}{[Pr(G, \lambda) - Pr(B, \lambda)] - [(Pr(G', \lambda) - Pr(B', \lambda]}$. Similarly $\beta\pi + (1 - \beta)\pi' \notin \hat{\Pi}$ for $\beta < \frac{-[Pr(G', \lambda) - Pr(B', \lambda)]}{[Pr(G, \lambda) - Pr(B, \lambda)] - [(Pr(G', \lambda) - Pr(B', \lambda]}$.

This concludes the proof. □

B.2.4 Extension: Principal has ex-ante correlation-sensitive preferences

In this section, I briefly discuss the case of a principal with ex-ante correlation-sensitive preferences. In the framework of Lanzani (2022), the crucial assumption made in regret and salience theory is increasing sensitivity to payoff differences (ISPD). That is, for all h, m, $l \in \mathbb{R}$, $h < m < l \Rightarrow \phi(h,l) > \phi(h,m) + \phi(m,l)$ (Loewenfeld and Zheng, 2023). For the lottery pairs considered in the experiment, this condition implies that the principal certainly prefers the dominant lottery when it yields a higher payoff only in one state but might prefer the dominant lottery when the dominant lottery yields a higher payoff in two states. When decreasing sensitivity to payoff differences (DSPD) holds, that is, $\phi(h,l) < \phi(h,m) + \phi(m,l)$, the reverse pattern might obtain.

Since OB is modeled as an ex-post increase of the salience of the state, the model can easily be adapted to account for correlation-sensitive preferences. Following the same logic as for the case in which the principal's preferences are described by expected utility theory, the principal awards the bonus after the agent chooses lottery θ and state s materializes whenever

$$bonus = 1 \iff \lambda_p \phi(x_s^\theta, x_s^{-\theta}) + (1-\lambda_p) \sum_{s=1}^{S} q_s \phi(x_s^\theta, x_s^{-\theta}) \tag{B.10}$$

Since $\phi(x_s^\theta, x_s^{-\theta}) > 0$ whenever $x_s^\theta > x_s^{-\theta}$, correlation sensitive preferences might change the magnitude, but not the sign of the ex-post counterfactual comparison. As a result, parts a) and b) of proposition 3 remain valid. That is, OB weakly reduces the incentives to choose the principal's preferred lottery (proposition part a)). When the principal's preferred lottery is less likely to yield a higher outcome than the alternative, a high enough level of OB can induce the principal to incentivize the agent to choose their least preferred lottery (proposition part b)).

Technically speaking, also part c) of the proposition remains valid, but it might not be an instant of the statement in part b). That is, a sufficiently outcome-biased principal might incentivize the agent to choose a dominated lottery, but this need not necessarily contradict her ex-ante preferences. If the principal satisfies ISPD, proposition b) applies, just as in the expected utility case. When a principal satisfies DSPD, the principal might have an ex-ante preference for a first-order stochastically dominated lottery when it is more likely to yield a higher outcome than the dominant lottery. If this is the case, proposition a) applies, but not b).

B.2.5 Comparison with existing models

Good decision with moral luck (Aimone and Pan, 2020): The model is similar to the one proposed here in the sense that reward decisions are based on the received quality of the agent's decision. Adapting notation to the current setting. the received quality is the sum of the 'pure quality' $f(\theta)$, which is determined by the principal's ex-ante preferences over the lotteries, and an outcome-based component $g(x_s^\theta)$, the moral luck component:

$$Receivedquality = f(\theta) + g(x_s^\theta) \tag{B.11}$$

Aimone and Pan (2020) consider a setting in which agents choose between a number of binary lotteries, each with a gain and a loss, and a safe option that yields a zero payoff with certainty. For this setting, they assume $g(x_s^\theta) = 1$ $(= -1)$ for $x_s^\theta > 0$ (< 0), and $g(x_s^\theta) = 0$ in case the safe option is chosen. This model could easily be adapted to the current setting with multiple possible outcomes. The main distinction between this model and the one proposed here is the absence of ex-post counterfactual comparison Aimone and Pan (2020)'s model. The model can, therefore, not account for any differences in the reward schemes of principals that arise from changes in the correlation of payoffs.

Salient perturbation of delegated expertise Gurdal et al. (2013): Gurdal et al. (2013) propose a model in which the principal behaves as if they were in a familiar setting that closely resembles the setting of choice between lotteries. They refer to this as the salient perturbation of the actual setting. They argue that the natural salient perturbation of the setting considered here (choices between lotteries) is a setting in which the agent can exert costly effort in order to better predict the resolution of uncertainty. In this parallel world, the agent can choose to exert costly effort in order to observe a costly and informative signal about the probability of different states of the world. If the agent exerts effort, he is more likely to choose a lottery that obtains a high outcome. Therefore, the principal's ex-ante optimal contract is to condition her bonus payments on the realized state of the world. In this fashion, the model can predict a dependence of reward decisions not only on obtained, but also forgone payoffs.

Although the idea of salient perturbations is very intriguing and the approach yields interesting insight, there are issues with this approach in the current setting, some of them mentioned by Gurdal et al. (2013). For instance, Gurdal et al. (2013) point out the following. The model predicts that, for a choice between a safe and a risky option, the reward paid to the agent if they choose the risky option and it yields a high payoff can be increased by lowering the probability of this payoff occurring (holding everything else constant). That is, holding the realized outcomes constant, the bonus payment in a given state can be increased by reducing the ex-ante attractiveness of the chosen lottery. In the setting of delegated expertise, this result obtains because the optimal contract needs to compensate the agent for a low ex-ante probability of the high state occurring, in case the agent observes a signal that makes it worth choosing the risky option. (Gurdal et al., 2013, p.1217) comment that whereas "this intuition is sensible in the delegated expertise setting, it is not in the experimental setting."

A further issue for the purposes of this paper is that the salient perturbation happens "in the head of the principal". Assumptions about the specifics of the salient perturbation, i.e., the relation of effort to signal and the structure of the signal observed by the agent, can generally influence conclusions, which makes it makes it difficult to use the salient perturbation approach to estimate principals' degree of outcome bias.

Inequity aversion: Models of inequity aversion (Fehr and Schmidt, 1999) predict that principals in the considered setting will condition their payoffs on the realized payoff, but not on the forgone payoff, and neither on the choice of the agent. This observation is generally true for models of other-regarding preferences that are defined over the final outcomes of individuals.

Reciprocity: Classical theories of reciprocity consider non-stochastic environments, such as the ultimatum game or the centipede game (Rabin, 1993; Falk and Fischbacher, 2006; Dufwenberg and Kirchsteiger, 2004). In these theories, a player's kindness is defined over their intention towards their opponent as measured by their actions. In settings in which an element of stochasticity is introduced, it has been noted that reciprocity is often also at least partly based on the realized outcome (Charness and Levine, 2007; Rubin and Sheremeta, 2016; Davis et al., 2017). However, none of these approaches considers the forgone outcomes.

Appendix C: Further Results

B.3.1 Non-parametric results

B.3.1.1 The principal's revealed preference

Averaged over all four lottery pairs, principals in the reward-after treatment chose lottery G_k^{1s} at a frequency of 91.7% and lottery G_k^{2s} at a frequency of 93.8% ($p = 0.362$, two-sided McNemar's exact test). Their counterparts in the reward-before treatment chose the dominant lotteries at 93.8% and 96.0%, respectively ($p = 0.286$, two-sided McNemar's exact test). This suggests that 1) principals do indeed perceive the dominant lotteries as superior, and 2) the identifying assumption that preferences are not significantly impacted by the change in correlation structure holds for the considered tasks. See also table B.1. Finally, treatment differences are small and statistically insignificant. The choice frequencies of G_k^{1s} were 91.7% in the reward-after vs 93.8% in the reward-before treatment ($p = 0.37$ two-sided, two-sample test of proportions). For the lotteries G_k^{3s}, the choice frequencies are 93.8% in the reward-after vs 96.0% in the reward-before treatment ($p = 0.25$).

	Reward-after				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
G_k^{1s}	87.7%	84.9%	97.3%	97.3 %	94.5%
G_k^{2s}	94.5%	91.8%	93.2%	95.9 %	-
$p-value$	0.267	0.180	0.375	1.000	-
	Reward-before				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
G_k^{1s}	91.2%	88.2%	98.5%	97.1%	97.1%
G_k^{2s}	92.6%	95.6%	97.1%	98.5%	-
$p-value$	1.000	0.125	1.000	1.000	-

Table notes: P-values are from exact McNemar tests.

Table B.1 The principals' lottery choices for themselves.

	Reward-after			
	(1)	(2)	(3)	(4)
	Pair 1	Pair 2	Pair 3	Pair 4
$Preferred$	0.369**	0.944***	1.236***	1.302***
	(0.182)	(0.183)	(0.205)	(0.225)
$obtained > forgone$	1.261***	0.911***	1.079***	1.005***
	(0.286)	(0.273)	(0.279)	(0.283)
$obtained\ payoff$ (in Euro)	0.129***	0.128***	0.189***	0.167***
	(0.020)	(0.020)	(0.029)	(0.028)
constant	-1.086***	-1.423***	-2.327***	-2.144***
	(0.191)	(0.192)	(0.224)	(0.222)
Observations	876	876	876	876
Individuals	73	73	73	73

	Reward-before			
	(1)	(2)	(3)	(4)
	Pair 1	Pair 2	Pair 3	Pair 4
$Preferred$	1.930***	2.112***	3.309***	3.166***
	(0.335)	(0.332)	(0.476)	(0.492)
$obtained > forgone$	0.305	-0.289	0.035	-0.757
	(0.531)	(0.492)	(0.591)	(0.592)
$obtained\ payoff$ (in Euro)	-0.001	0.034	0.114**	0.074
	(0.035)	(0.033)	(0.057)	(0.054)
constant	-0.303	-0.509*	-2.148***	-1.386***
	(0.266)	(0.273)	(0.415)	(0.339)
Observations	272	272	272	272
Individuals	68	68	68	68

Table notes: Standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1.

Table B.2 Coefficients of the logistic model.

Lottery	Reward-before				Reward-after				diff-in-diff
pair	$I(G^{1s})$	$I(G^{2s})$	$I(G^{1s})$- $I(G^{2s})$	p-value	$I(G^{1s})$	$I(G^{2s})$	$I(G^{1s})$ - $I(G^{2s})$	p-value	p-value
1	0.279	0.412	-0.133	0.293	-0.055	0.136	-0.191	< 0.001	0.064
2	0.426	0.426	0	0.881	0.110	0.192	-0.082	0.046	0.085
3	0.632	0.632	0	1.000	0.215	0.365	-0.150	0.001	0.002
4	0.632	0.706	-0.074	0.201	0.306	0.343	-0.037	0.070	0.893

Table notes: The last column "diff-in diff" refers to the p-values when testing the null $I(G_k^{1s}, before) - I(G_k^{2s}, before) = I(G_k^{1s}, after) - I(G_k^{2s}, after)$. As preregistered, p-values from one-sided Wilcoxon rank-sum tests are reported for this comparison. As no directional alternative was preregistered for the other test for which p-values are reported here, these are from two-sided Wilcoxon signed-rank tests.

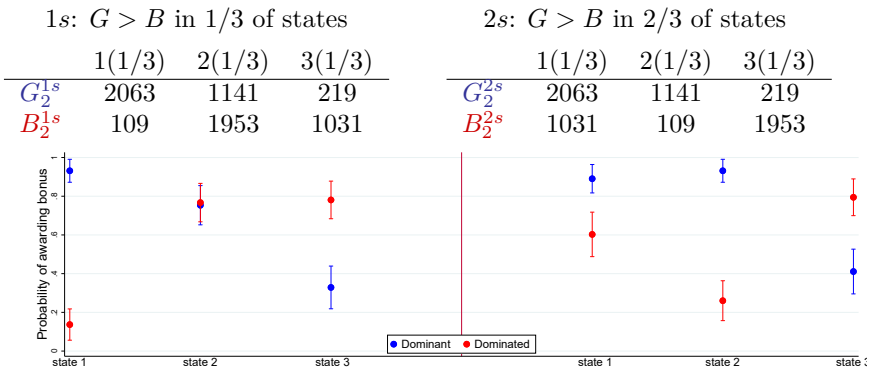
Table B.3 Incentives under different correlation structures.

B.3.1.2 Supplementary tables and figures

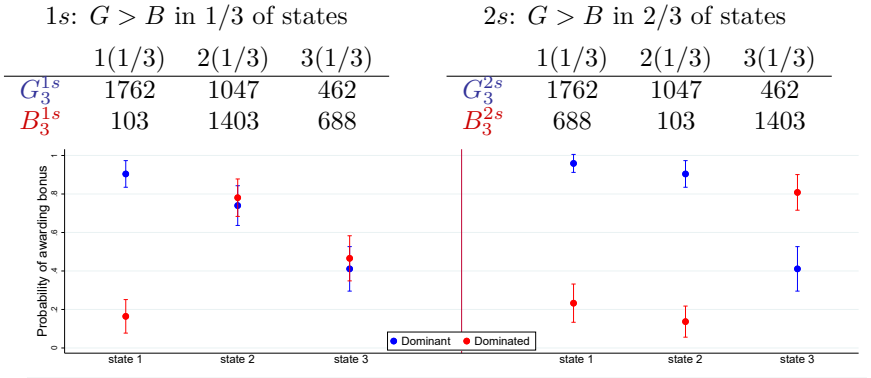
Lottery	Reward-before				Reward-after				diff-in-diff
pair	$F(G^{1s})$	$F(G^{2s})$	$F(G^{1s})$- $F(G^{2s})$	p-value	$F(G^{1s})$	$F(G^{2s})$	$F(G)$ - $F(G')$	p-value	p-value
1	75.0%	87.5%	-12.5ppts	0.019	70.5%	94.5%	-24.0ppts	< 0.001	0.043
2	81.6%	92.6%	-11.0ppts	0.003	68.5%	93.2%	-24.7ppts	< 0.001	0.027
3	86.8%	91.9%	-5.1ppts	0.257	87.0%	97.9%	-10.9ppts	0.002	0.107
4	89.0%	95.6%	-6.6ppts	0.021	91.8%	96.6%	-4.8ppts	0.184	0.730

Table notes: The last column "diff-in diff" refers to the p-values when testing the null $F(G_k^{1s}, before) - F(G_k^{2s}, before) = F(G_k^{1s}, after) - F(G_k^{2s}, after)$. As preregistered, p-values from one-sided Wilcoxon rank-sum tests are reported for this comparison. As no directional alternative was preregistered for the other test for which p-values are reported here, these are from two-sided Wilcoxon signed-rank tests.

Table B.4 Incentives under different correlation structures.



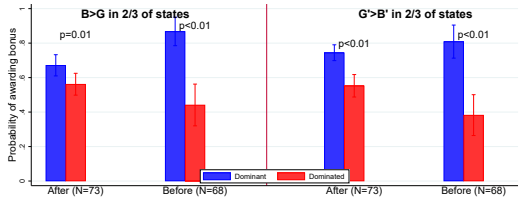
(a) Lottery pair 2



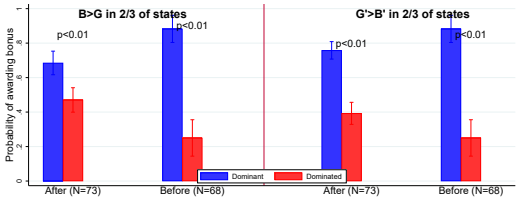
(b) Lottery pair 3

Table notes: Award frequencies are averaged over the 73 principals in the reward-after treatment, with 95% confidence intervals. The states are indicated on the x-axis and are below the corresponding state in the table displaying the choice task.

Figure B.1 Frequencies of awarding the bonus for a given choice-state combination.



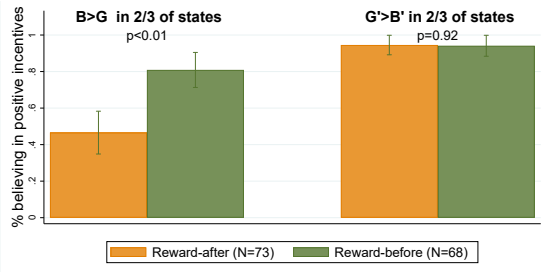
(a) Lottery pair 2



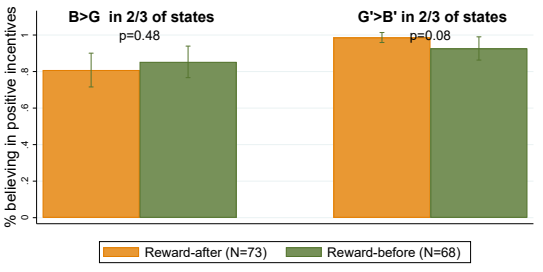
(b) Lottery pair 3

Notes: With 95% confidence intervals. For the reward-after treatment, the frequencies are averaged over the three states.

Figure B.2 Frequencies of awarding the bonus, conditional on the agent's lottery choice.

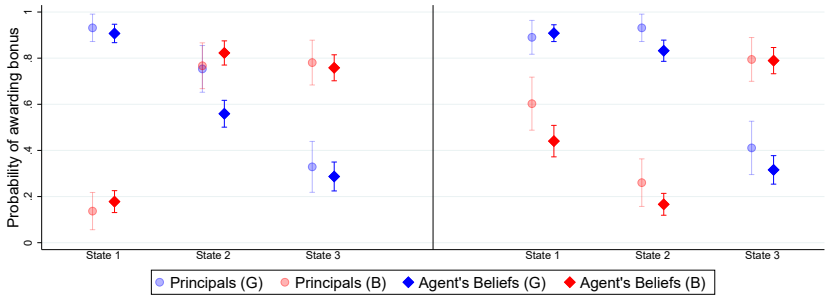


(a) Lottery pair 2

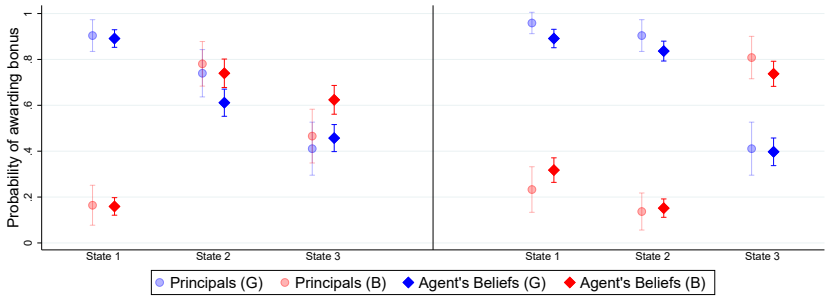


(b) Lottery pair 3

Figure B.4 The proportion of agents who believe to have weakly positive incentives to choose the dominant lottery, with 95% confidence intervals.



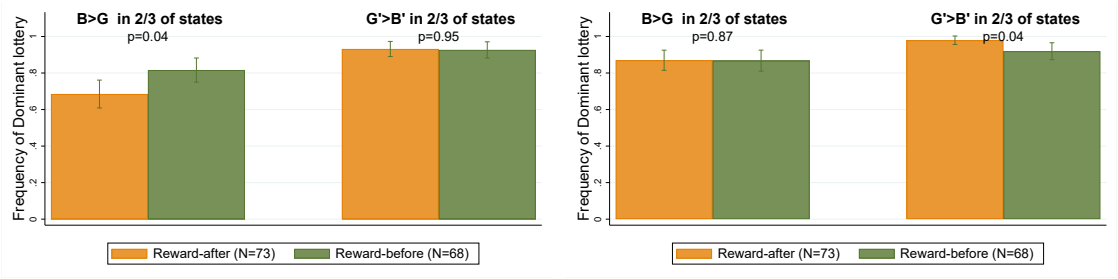
(a) Lottery pair 2



(b) Lottery pair 3

Notes: With 95%. For reference, the principals' bonus decisions are displayed in transparent circles.

Figure B.3 Average beliefs of the 73 agents in the reward-after treatment, conditional on the lottery choice and the realized state of the world.



(a) Lottery pair 2 **(b)** Lottery pair 3

Notes: The choices are averaged over the first-second choice agents make for each lottery task, with 95% confidence intervals.

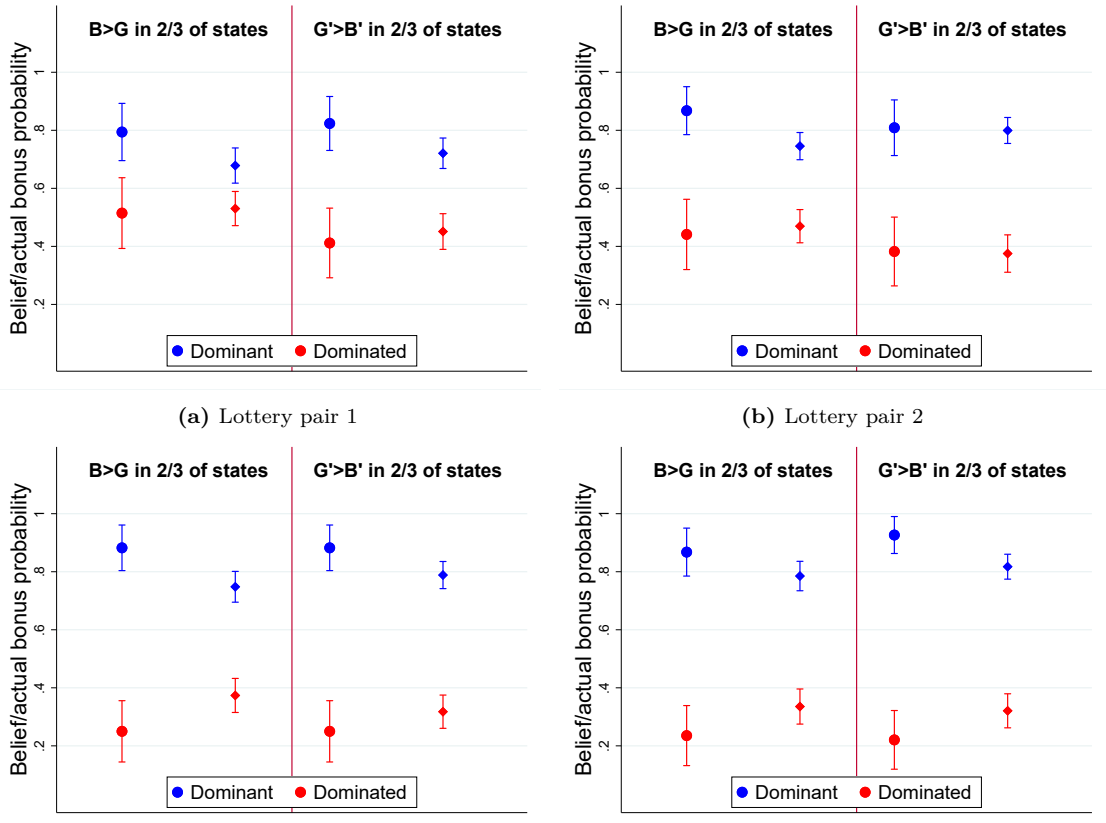
Figure B.5 The frequency with which agents choose the dominant lottery.

B.3.1.3 Further details on beliefs

Systematic deviation of beliefs from actual bonus decisions: Although agents in both treatments anticipate the broad patterns of the principal's bonus decisions, beliefs deviate from bonus decisions in a systematic way. Averaging over all choice-state combinations, agents in both treatments somewhat underestimate the probability of receiving the bonus when choosing the dominant lottery. In the reward-before (reward-after) treatment, the average belief is 67.5% (77.0%), whereas principals award the bonus with 72.6% (85.9%) probability ($p = 0.045$,[9] in the reward-after and $p = 0.004$ in the reward-after treatment). Beliefs about the probability of receiving the bonus when choosing the dominated lottery do not differ significantly from the actual bonus probabil-

[9] P-values in this and paragraph are obtained from two-sided t-tests, clustered at the individual level to control for repeated observations from the same individual.

ity.[10] Overall, agents in the reward-before treatment significantly underestimate their incentives to choose the dominant lottery by 14.6ppts (47.3ppts (beliefs) vs 66.9ppts (actual incentives), $p < 0.001$), whereas their counterparts in the reward-after treatment do not significantly over- or underestimate their incentives(18.2ppts (beliefs) vs. 21.8ppts (actual incentives), $p = 0.319$).



(a) Lottery pair 1 **(b)** Lottery pair 2

(c) Lottery pair 3 **(d)** Lottery pair 4

Figure B.6 The agents' beliefs in the reward-before treatment

B.3.1.4 Further details on anticipated incentives

First, in the reward-after treatment, a lower proportion of agents believe to have weakly positive incentives to choose the dominant lottery when it yields a higher payoff in only 1/3 of states than when it yields a higher payoff in most states. The differences are statistically significant for lottery pairs 1, 2, and 3 ($p < 0.002$, two-sided exact McNemar test) and statistically insignificant for lottery pair 4 ($p = 0.227$). Especially for lottery pairs 1 and 2, the differences are sizable. They amount to roughly 42ppts for lottery pair 1 and 48ppts for lottery pair 2. Agents in the reward-before treatment display somewhat of a similar tendency, but the magnitude is much smaller and statistically insignificant safe for lottery pair 2.[11].

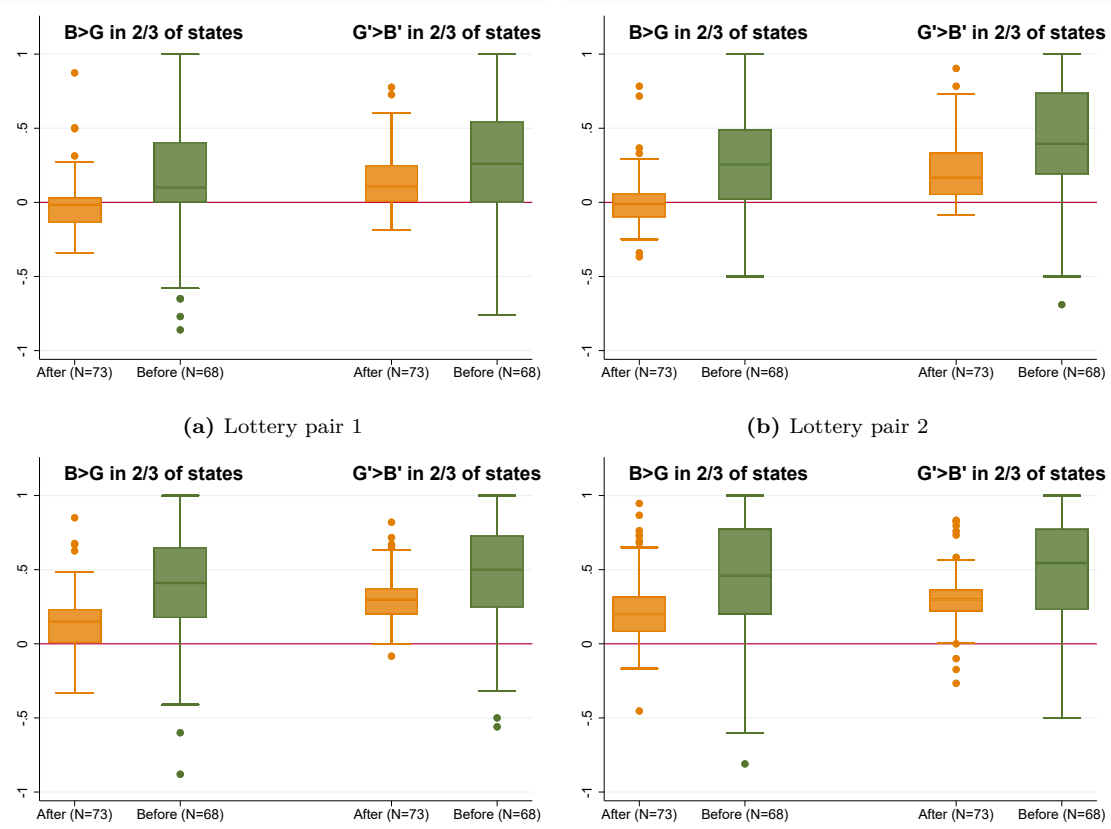
Second, drawing comparisons across treatments, agents in the reward-after treatment are significantly less likely to believe to have positive incentives to choose the dominant lottery when it

[10] In the reward-after (reward-before) treatment, agents receive the bonus with 50.7% (32.7%) probability when choosing the dominated lottery, and the average belief is 49.3% (38.4%) ($p = 0.598$ in the reward-after, and $p = 0.165$ in the reward-before treatment).

[11] Lottery pair 1: 86.7% vs. 76.5% ($p = 0.119$); lottery pair 2: 94.1% vs. 80.8% ($p = 0.035$); lottery pair 3: 92.6% vs. 85.3% ($p = 0.267$); lottery pair 4: 94.1% vs. 88.2% ($p = 0.344$). P-values are from two-sided exact McNemar test.

yields a higher payoff in just one state for lottery pairs 1 and 2, but not lottery pair 3 and 4.[12] The treatment differences for lottery pairs 1 and 2 are quite sizable and amount to roughly 34ppts for both lottery pair 1 and 2. No significant treatment differences arise when the dominant lottery yields a higher payoff in 2/3 of the states.[13]

Figure B.7 shows boxplots of the agents' beliefs regarding their incentives to choose the dominant lottery for the four lottery pairs and both treatments. The boxplots are useful to display the considerable variance in agents' beliefs. As self-interested agents should choose the lottery that will provide a higher expected probability of receiving the bonus, this variance is important to understand. Consider first the beliefs of agents in the reward-after treatment. Within each lottery pair, beliefs about the incentives to choose lottery G' are shifted upwards relative to the beliefs about the incentives to choose lottery G. Further, across the four lottery pairs, the beliefs about the incentives to choose the dominant lottery are increasing from pair 1 to 2, 3, and 4.



(a) Lottery pair 1

(b) Lottery pair 2

(c) Lottery pair 3

(d) Lottery pair 4

Figure B.7 The agents' beliefs in the reward-after treatment

For lottery pair 1, when it is the dominant lottery G yields a lower payoff in 2/3 states, a majority of 58% of the agents in the reward-after treatment believe to be more likely to obtain the bonus when choosing the dominant rather than the dominated lottery. The agent at the 25percentile believes to be 14 points more likely to choose the obtain the bonus when doing so. However, when

[12] Lottery pair 1: 42.5% (reward-before) vs. 76.5% (reward-after)% ($p < 0.001$); lottery pair 2: 46.6% vs. 80.8% ($p < 0.001$); lottery pair 3: 80.8% vs. 85.3% ($p = 0.480$); lottery pair 4: 89.0% vs. 88.2% ($p = 0.880$). P-values are from two-sample tests of proportions.

[13] Lottery pair 1: 84.9% (reward-before) vs. 86.8% (reward-after)% ($p = 0.755$); lottery pair 2: 94.5% vs. 94.5% ($p = 0.918$); lottery pair 3: 98.6% vs. 92.6% ($p = 0.079$); lottery pair 4: 95.9% vs. 94.1% ($p = 0.628$). P-values are from two-sample tests of proportions.

the dominant lottery lottery G' yields a higher payoff in 2/3 of the states, a majority of 85% of the agents belief to be weakly more likely to obtain the bonus when choosing the dominant lottery. For lottery pair 2, the boxplots are shifted upwards, but there is still a small majority of 53% of the agents in the reward-after treatment believe to be strictly more likely to obtain the bonus when choosing the dominated lottery B, although principals are 11ppts more likely to award the bonus for the dominant lottery. A large majority of 95% of the agents believe to be weakly more likely to receive a bonus when choosing the dominant lottery G'. For lottery pairs 3 and 4, the overwhelming majority of the agents believe to have positive incentives to choose the dominant lottery.

Consider now the beliefs about incentives of agents in the reward-before treatment. For each lottery pair, beliefs for the incentives to choose lottery G and G' do not seem to differ much. Going from lottery task 1 to 2, 3, and 4, the boxplots are shifted up, although a vast majority of the agents believe to have positive incentives to choose the lotteries for all lottery pairs. The boxplots for the reward-before treatment have a larger range than those for the reward-before treatment. This is likely due to the fact that the beliefs over incentives are calculated from 2 stated beliefs in the reward-before, but six stated beliefs in the reward-after condition, which allows for more averaging out of noise in the reward-before treatment.

B.3.2 Details on inconsistency between stated beliefs and choices

Rates of Consistency: I define a lottery choice to be consistent with an agent's beliefs if the agent believes to have weak incentives to choose the dominant lottery and chooses the dominant lottery or believes to have strictly negative incentives to choose the dominant lottery and chooses the dominated one. Otherwise, a choice is said to be inconsistent. According to the classification, in the reward-after treatment for lottery pair 1 and 2 under 1-state-better correlation structure, only 55% and 57% of the lottery choices are consistent with beliefs. Neither of these proportions differs significantly from the 50% benchmark that could be achieved by random choice ($p = 0.483$ and $p = 0.242$, two-sided binomial test). Inconsistency is especially stark among agents whose beliefs indicate strict incentives to choose the dominated lottery. Among the 42 (39) agents in the reward-after treatment who believe to be strictly more likely to obtain the bonus when choosing the dominated lottery, a majority of 61.9% (61.5%) chose the dominant lottery anyhow. Note that this suggests that noisy best-response to beliefs alone is unlikely to explain the inconsistency between beliefs and choices. Noisy best-response suggests an upper bound of 50% dominant lottery choices, but the frequencies are significantly larger than 50% at $p = 0.021$ for lottery pair 1 and $p = 0.089$ for lottery pair 2 (two-sided binomial test). Moreover, the rates of consistency range between 83.5% and 97.2% for the other choice tasks, all of which are significantly different from 50% at $p < 0.001$. Likewise, consistency rates in the reward-before treatment range between 79.4% and 91.2%, all of which are significantly different from 50% at $p < 0.001$.

Predictability of choices for beliefs: Table B.5 shows results from regressions exercises that probe how predictive beliefs are of choices. To simplify the analysis, I focus on the agents' second choice. As they make this choice directly after the belief elicitation, one might expect beliefs

to be especially predictive of choices for agents' second choice. I estimate the following random effect-logistic model.

$$G_{corr,k,a} = \beta_0 + \beta_1 positive_{corr,k,a} + \beta_2 before_a + \beta_3 positive_{corr,k,a} * before_a + \epsilon_{corr,k,a}, \quad \text{(B.12)}$$

where $G_{corr,k,a}$ is a dummy that is equal to one if an agent a chose the dominant lottery for a given choice task and zero otherwise. $positive_{corr,k,a}$ equals one if an agent believes to have weakly positive incentives to choose the dominant lottery for a given choice task, and $before$ is a dummy that indicates whether a given agent completed the experiment in the reward-before treatment. For convenience of interpretation, table B.5 shows both the estimated coefficients and AMEs.

	Sample 1		Sample 2	
	(1)	(2)	(3)	(4)
	Logistic	AMEs	Logistic	AMEs
VARIABLES	$G_{corr,k,a}$	$G_{corr,k,a}$	$G_{corr,k,a}$	$G_{corr,k,a}$
positive	0.313	0.022	3.052***	0.150***
	(0.913)	(0.064)	(0.627)	(0.032)
before	-4.208***	-0.295***	-0.136	-0.007
	(1.566)	(0.087)	(0.734)	(0.036)
positive $*$ *before*	7.723***	0.540***	-0.679	-0.033
	(2.332)	(0.118)	(0.815)	(0.040)
Constant	2.035**		1.373***	
	(0.853)		(0.525)	
Observations	282	282	846	846
Number of agents	141	141	141	141

Table notes: Columns (1) and (3) report the estimated coefficients, and Columns (2) and (4) report AMEs. Sample 1 refers to the choice tasks of lottery pairs 1 and 2 for which the dominant lottery yields a higher payoff in just one state. Sample 2 refer to the remaining choice tasks. Standard errors in parentheses.*** p<0.01, ** p<0.05, * p<0.1.

Table B.5 Regression analysis probing how predictive beliefs are about choices.

Sample 1 in columns (1) and (2) zooms in on the two choice tasks for which agents' actions seem inconsistent with their choices, that is, the correlation structure when the dominant lottery yields a higher payoff in one state only for lottery pairs 1 and 2. As figure 4.4 illustrates, we see the greatest variance in beliefs regarding incentives for these two choice tasks and the reward-after treatment. As such, one might expect beliefs to be especially predictive of choices for agents in the reward-before treatment. However, in the reward-after treatment, the AME effect of believing to have incentives to choose the dominant lottery is estimated at a mere 1.6ppts and is not statistically different from zero ($p = 0.807$). Relative to their peers in the reward-after treatment, agents in the reward-before treatment are estimated to be 29.5 ppts less likely to choose the dominant lottery when they believe to negative incentives to do so ($p = 0.001$). They are also 24.7 ppts more likely than agents in the reward-after treatment to choose the dominant lottery when they believe they have weakly positive incentives to do so. However, this is not statistically significant ($p = 0.114$, chi-square test of the null hypothesis that $\beta_1 = \beta_2 + \beta_3$).

Sample 2 in columns (3) and (4) shows regression results for the remaining choice tasks, for which choices are largely consistent with stated beliefs. The AMEs indicate that agents in the reward-after treatment who believe to have positive incentives to choose the dominant lottery are 15 ppts more likely to choose the dominant lottery ($p < 0.001$). The estimated coefficients for β_2 and β_3 are not statistically different from zero, and the corresponding AMEs are close to zero. This suggests that the beliefs of agents in the two treatments are similarly predictive of choices. Although beliefs are found to be predictive of choices, the effect size might be regarded as small. This lends further support to the notion that agents might employ two largely different modes of reason when stating their beliefs and when choosing between the lotteries.

Response time: Exploratory analysis suggests that the inconsistency between the agents' beliefs and their choices might be due to a failure of agents to even consider their beliefs before making their lottery choice. Recall that agents made a first lottery choice for each lottery task, which was then followed by a second lottery choice. Averaging over all four lottery pairs, agents in the reward-after treatment take 15.3 seconds for their first lottery choice. When asked to state their beliefs, they take an average of 55.12 seconds to enter their beliefs for all six choice-state combinations of a lottery task (3 (states) $\times$ 2(choices)). If agents were to form beliefs about the principals' bonus decisions before making their first lottery choice, one would expect agents to take much longer for their first choice. Upon being asked to enter their beliefs, agents, having already thought about them, should be quick to enter them. It seems unlikely that the large discrepancy between the response times for beliefs and choices results simply from the fact that agents have to enter six different beliefs but have to make only one choice. Likewise, one might argue that forming beliefs about incentives is more challenging in the reward-after as subjects in the reward-after treatment have to consider all six possible choice-state combinations, whereas subjects in the reward-before treatment only have to consider the principal's bonus decisions conditional on their choices. If agents were to form beliefs before making their first choice, one might expect agents in the reward-after condition to take more time to make this choice. However, the average response time is virtually the same in both treatments (15.3 (reward-after sec vs 14.9 sec. $p = 0.760$ two-sided t-test, standard errors clustered at the individual level).

Distributional preferences: Could other-regarding preferences explain the results? This also seems unlikely, as standard models of other-regarding preferences fail to rationalize the observed choice patterns. First, if an agent's preferences are described by inequity aversion (Fehr and Schmidt, 1999), holding the bonus probability fixed, for lottery pairs 1 and 2, a behindness averse agent would rather choose the dominated lottery. To see this, note that the high and the medium payoff of the lottery are larger than the bonus of 10 Euros an agent can obtain. For these two realizations, the agent's payoff will be lower than that of the principal, regardless of whether he obtains the bonus or not. He would rather not have the principal enjoy the extra premium that comes with the dominant lottery. For the third realization, the agent would rather have the principal enjoy the premium of the dominant lottery if he receives the bonus but not if he doesn't. Overall, a behindness averse agent, therefore, prefers the principal to be paid according

to the dominated lottery.[14] Second, a model of pure altruism (Becker) does not seem to be able to explain the agent's choices. For lottery pair 1, the premium of the dominant lottery is small compared to the size of the bonus payment. One can show that, in order to choose the dominant lottery despite believing to have incentives to choose the dominated lottery, a large number of agents would be required to care more about the principals' than their own payoff, which seems unrealistic. More generally, any model in which the agent trades off his own payoff with that of the principal would still imply that beliefs are predictive of choices when there is a conflict between the agent's and the principal's payoff.

	Reward-after		Reward-before	
	(1)	(2)	(3)	(4)
VARIABLES	Consistent	Consistent	Consistent	Consistent
positive	0.300**	0.437***	0.305*	0.370***
	(0.149)	(0.112)	(0.159)	(0.115)
CRT	-0.037		-0.039	
	(0.028)		(0.043)	
$CRT * positive$	0.034		0.047	
	(0.037)		(0.046)	
$Econ$		-0.015		-0.120
		(0.113)		(0.160)
$Econ * pos$		-0.014		0.164
		(0.156)		(0.179)
Constant	0.519***	0.385***	0.487***	0.431***
	(0.122)	(0.084)	(0.141)	(0.102)
Observations	146	146	136	136
Number of subjects	73	73	68	68

Table notes: The dependent variable consistent is equal to 1 if an agent's stated beliefs indicate weakly positive incentives to choose the dominant lottery and the agent chooses the dominant lottery or if beliefs indicate strictly positive incentives to choose the dominated lottery and the agent chooses the dominated lottery. Otherwise, the variable equals 0. The sample consists of the set of first choices agents make before the belief elicitation. The table displays coefficients of random-effects regressions. Standard errors in parentheses.*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.6 Consistency of choices and stated beliefs - first round.

Appendix D: Details on the estimation

B.4.1 Individual level - Principals

Principal p states her bonus decisions conditional on the realized state $s \in \{1, 2, 3\}$ and the agent choosing a lottery θ_l. To keep notation manageable, I denote each of the nine lottery tasks employed in the principal-agent interaction by subscript $l \in \{1, ..., 9\}$. To derive choice probabilities within a random utility framework, I consider the utility function of a principal whose bonus decisions are motivated by outcome-biased reciprocity, as derived in section B.2.1. The utility function of principal p, after observing lottery choice θ_l and state s is given by

[14] In the notation of Fehr and Schmidt (1999), the agent would prefer the principal to receive the dominant lottery only if $\beta > 2\alpha$, which contradicts the assumption that $\beta \leq \alpha$.

$$u_p(bon_{p,\theta_l,s}, \rho_p, \lambda_p, r_p) = \\ u(x_s^{\theta_l}) + 0.5\rho(bon_{p,\theta_l,s} - 0.5)[\lambda(u(x_s^{\theta_l}) - u(x_s^{-\theta_l})) + (1-\lambda)(EU^{\theta_l} - EU^{-\theta_l}] \tag{B.13}$$

The principal maximizes her utility function by choosing whether to give the bonus to the agent or not. When taking a deterministic model to a setting with noisy decision-making, one has to make assumptions about the shape of errors. I employ a standard random utility model. Principal p calculates her utility subject to an additive error, that is

$$\tilde{u}_p(bon_{p,\theta_l,s}, \rho_p, \lambda_p, r_p) = u_p(bon_{p,\theta_l,s}, \rho_p, \lambda_p, r_p) + \epsilon_{bon_{p,\theta_l,s}}, \tag{B.14}$$

where $\epsilon_{p,\theta_k,s,bonus_{p,\theta_k,s}}$ is assumed to be i.i.d. and distributed according to the type 1 extreme value distribution. The principal decides to award the bonus whenever the random utility of awarding the bonus is higher than the random utility of not awarding it. It is a well-known result that, under the chosen error structure, choice probabilities follow the logit formula (McFadden, 1973). Therefore the probability of that principal p chooses $bon_{p,\theta_k,s} = 1$, denoted by $pr(\lambda_p, r_p, \mu_p^1, \theta_k^{corr}, s)$, is given by expression B.27, with $\mu_p^1 = \rho_p \mu_p$, where μ_p is a noise parameter. As we will see shortly, the reciprocity parameter ρ_p is only identified under strong assumptions, which is why I prefer to use a generic noise parameter μ_p^1.

Define $bon(p, \theta_l, s) \in \{0, 1\}$ an indicator function that equals 1 if a principal awarded the bonus for a given lottery choice and state, and 0 otherwise. Given expression B.27, the probability density function for bonus decisions can be written as

$$f^{bonus}(\lambda_p, r_p, \mu_p^1) = \\ \prod_{l \in \{1,...9\}} \prod_{s \in S} \Big(bon(p, \theta_l, s) pr(\lambda_p, r_p, \mu_p^1, \theta_l, s) + (1 - bon(p, \theta_l, s))(1 - pr(\lambda_p, r_p, \mu_p^1, \theta_l, s)) \Big) \tag{B.15}$$

I estimate the two parameters (λ_p, μ_p^1) via MLE.

B.4.1.1 Allowing for risk preferences

In the main specification, I impose a linear utility function. Here, I discuss the estimation procedure when allowing for risk preferences a parameterized by CRRA utility function, that is $u(x) = \frac{x^{1-r_p}}{1-r_p}$ for $r_p \neq 1$ and $u(x) = log(x)$ for $r_p = 1$. Bonus decisions alone do not provide sufficient variation to identify the parameter of utility curvature r_p. Therefore, the parameter is mainly identified from lottery choices principals make for themselves at the end of the experiment. These lottery choices include the nine choice tasks used in the principal-agent interaction, as well as three multiple-choice lists totaling 33 binary choices. That is, principals had to select a lottery $\theta_l \in \{A_l, B_l\}$, for $l \in \{1, ..., 42\}$. Applying again a random utility model and assuming an i.i.d. type-1 extreme value distribution of errors, the probability of selecting lottery A_l is given by

$$pr(r_p, \mu_p^R, l) = \frac{1}{1 + exp(-\mu_p^R(EU(A_l) - EU(B_l))}, \tag{B.16}$$

where μ_p^R is a noise parameter governing the principal's lottery choices. We can now see that it would be possible to identify a principal's parameter of reciprocity ρ_p from how noisy bonus decisions are implemented relative to lottery choices, by assuming that the noise parameter is the same for lottery choices and bonus decision, that is, imposing $\mu_p^R = \mu$. However, the identifying assumption seems very strong, and estimating the reciprocity parameter is not the concern of this paper. Therefore, I prefer to simply refer to the more generic noise parameters μ_p^1 and μ_p^R.

Define $I(p, l)$ an indicator function that equals 1 if principals p chose lottery A for lottery task l. The probability density function for a principal's lottery choices can then be written as

$$f^{risk}(r_p, \mu_p^R) = \prod_l \Big(I(p,l) pr(r_p, \mu_p^R, l) + (1 - I(p,l))(1 - pr(r_p, \mu_p^R, l)) \Big) \tag{B.17}$$

As pointed out by (Apesteguia and Ballester, 2018) using a random utility model to estimate risk preferences can be problematic because choice probabilities are not monotonic for some range of preference parameters. In the estimation, I normalize utilities such that $u(x) \in [0, 1]$. This solves the issue for the relevant range of CRRA parameters. Without the normalization, I indeed find that choice probabilities are non-monotonic and simulation exercises show that the CRRA parameter can not be recovered well with the RUM approach, as pointed out by Apesteguia and Ballester (2018). With the normalization, choice probabilities are monotonous for the relevant range of CRRA parameters, i.e. $r \in (-1, 1)$. Simulation exercises show that the CRRA parameters can be recovered well without bias. Beyond $r \in (-1, 1)$, the identification is not very good but this is not very surprising given that the lottery tasks used to elicit risk preferences so not cover much more than this range.[15]

I estimate the four parameters $(\lambda_p, \mu_p^1, r1, p, \mu_p^R)$ jointly, using maximum likelihood estimation. Assuming independence of error terms, the joint probability density function of bonus decisions and lottery choices be written as

$$f^P(\lambda_p, r_p, \mu_p^1, \mu_p^R) = f^{bonus}(\lambda_p, r_p, \mu_p^1) f^{risk}(r_p, \mu_p^R) \tag{B.18}$$

B.4.2 Individual level - Agents

For each possible lottery choice θ_l and realized state s, agent a is asked to state the probability with which they expect to receive the bonus from a randomly chosen principal. I assume that agents arrive at these beliefs by forming beliefs over the principal's preference parameters. To keep the estimation tractable, I will mainly assume that the agent forms their beliefs by thinking of a representative principal, meaning that the belief about the bonus probability simplifies to, That is, agents calculate the probability of obtaining a bonus as follows

[15] As Apesteguia and Ballester (2018) point out, normalization of utilities does not generally solve the issue of non-monotonic choice probabilities. It is possible that the normalization simply moves the problematic range of parameter to one which is without much practical importance.

$$E_a[pr(bonus_{\theta_k,s})] = pr(bonus_{p,\theta_l,s}, \check{\lambda}_a, \check{r}_a, \check{\mu}_a^1), \tag{B.19}$$

where $pr(bonus_{p,\theta_l,s}, \check{\lambda}_a, \check{r}_a, \check{\mu}_a^1)$ is derived from the logit model as specified in equation B.27. I assume that agents implement these beliefs with an error, that is, they state[16]

$$\tilde{E}_a[pr(bonus_{\theta_l,s})] = E_a[pr(bonus_{\theta_l,s})] + \epsilon_{a,\theta_l,s}, \tag{B.20}$$

where $\epsilon_{a,\theta_l,s}$ is assumed to be a mean zero i.i.d. error distributed according to $N(0, 1/\mu_2)$. Note that this specification implies that reported beliefs are censored at 0 and 1, because agents are required to report a probability. In line with this formulation, a significant fraction of 9.9% of the stated beliefs are equal to zero or 1.

The idea behind the maximum likelihood estimation is to minimize the distance between the agents' stated beliefs and the beliefs produced by the model, $E_a[pr(bonus_{\theta_k,s})]$. Denote $SB(\theta_l, s, a) \in [0, 1]$ the stated belief. The probability distribution function of $\big(E_a[pr(bonus_{\theta_l,s})] - SB(\theta_l, s, a)\big) \in [0, 1]$ follows a censored normal given by

$$f^A(\check{\lambda}_a, \check{r}_a, \check{\mu}_a^1, \mu_a^3, \theta_l, s) = \begin{cases} \Phi(-SB(\theta_l, s, a))\mu_a^2), & \text{if } SB(\theta_l, s, a) = 0 \\ \mu_a^3 \phi\big(E_a[pr(\check{\lambda}_a, \check{r}_a, \check{\mu}_a^1, \theta_l, s)] - SB(\theta_l, s, a)\big), & \text{if } SB(\theta_l, s, a) \in (0, 1) \\ \Phi(SB(\theta_l, s, a)/\mu_a^2), & \text{if } SB(\theta_l, s, a) = 1 \end{cases},$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the distribution function and the probability distribution function of a standard normal, respectively. I then precede by maximum likelihood estimation.

B.4.3 Finite mixture models

I estimate finite mixture models (FMM) that assume a given number of preference types, similar to approaches taken to estimate risk (Bruhin et al., 2010b) or social preferences (Bruhin et al., 2019; Van Leeuwen and Alger, 2019). An advantage of finite mixture models is that they allow capturing heterogeneity while maintaining a high level of parsimony. Rather than estimating a set of parameters for each of N individuals, the finite mixture model assumes that there are G distinct preference classes. For a role $J \in \{P, A\}$ (and associated subscript $r \in \{p, a\}$), denote the ex-ante probability of membership in a given class g by γ_g^J. Further, denote $\Psi_g^J \in \{(\lambda_g, r_g), (\check{\lambda}_g, \check{\mu}_g^1, \mu_g^2\}$ the set of preference parameters to be estimated for a class g on the principals' and the agents' side. The probability density functions can be written as (McLachlan et al., 2019)

[16] An alternative specification would be to formulate a logit model, leveraging the incentive structure implied by the binarized scoring rule. However, as instructions were provided to subjects only after they clicked on a link, following the procedure recommended in Danz et al. (2022), most subjects were not aware of the actual rule, but only that they would be more likely to receive a reward the more accurate their stated belief is. I therefore consider the specification discussed here more appropriate.

$$\prod_{r=1}^{J}\sum_{g=1}^{G}\gamma_g^R f^J(\Psi_g^J) \tag{B.21}$$

Given parameters Ψ_g^J and γ_g^J for each individual r, one can calculate the posterior probability that an individual r belongs to a given class g according to Bayes rule as (McLachlan et al., 2019)

$$\tau_g^j = \frac{\gamma_g^J f(\Psi_g^J)}{\sum_{g=1}^{G}\gamma_g^J f(\Psi_g^J)} \tag{B.22}$$

The posterior probabilities provide a measure of individual preferences and also allow gauging how precisely individual participants are assigned to the different classes. Posteriors close to zero and one indicated precise assignment, whereas intermediate probabilities reflect a lack thereof.

B.4.4 Selection between different models

While allowing for additional classes increases the fit of the estimation, it leads to a less parsimonious typology and might reduce the accuracy with which individuals are assigned to different classes. To select between different models, a number of criteria exist. McLachlan et al. (2019) recommends the integrated classification (ICL) criterion developed by Biernacki et al. (2000). The ICL is given by

$$ICL = -2logL(G) + dlogN + EN(G), \tag{B.23}$$

where $logL(G)$ is the log-likelihood of the model with G classes, d is the number of estimated parameters, N is the number of individuals in the sample, and the entropy $EN(G)$ is a measure of how precisely individuals are assigned to different types for the model with G types. The entropy criterion is given by

$$EN(G) = -\sum_{g-1}^{G}\sum_{r=1}^{R}\hat{\tau}_g^j log(\hat{\tau}_g^j) \tag{B.24}$$

Note that the entropy is maximized when individuals are assigned to different classes with probabilities approaching zero or one. The ICL is equivalent to the BIC but adds an extra term that penalizes entropy. Biernacki et al. (2000) run simulation exercises and find that the ICL often outperforms the BIC in selecting the correct number of classes. As with the BIC, smaller values indicate more support for a given model.

I further follow Bruhin et al. (2010b) Bruhin et al. (2019) and Van Leeuwen and Alger (2019) in also considering the normalized entropy criterion (NEC) proposed by Celeux and Soromenho (1996), which is defined as

$$NEC(G) = \frac{EN(G)}{logL(G) - logL(1)}, \tag{B.25}$$

where $logL(1)$ is the log-likelihood of the one class model, and $logL(G)$ is that of a model of G classes. The NEC thus normalizes the entropy of a model by the gain in fit relative to the one-class model. Note that the NEC is not defined for the one-class model. It can, therefore, only be used to select between models with more than one class. The NEC also selects the model with the lowest value.

B.4.5 Choice tasks and Identification

The principals' bonus decisions and the agents' beliefs are elicited for a total of 54 choice-state combinations. In addition, principals make a total of 42 lottery choices. Simulation exercises suggest that all parameters are well identified. In the following, I provide an intuition for how the parameters are identified.

In the deterministic model introduced in section 4.2, a threshold value of $\underline{\lambda} = \frac{\Delta(EU)}{\Delta(EU)-\Delta_s}$ can be calculated such that principals with $\lambda \in [\underline{\lambda}, 1]$ will award the bonus. The parameters of the choice tasks are chosen to cover a wide range of cut-off values. For instance, assuming linear utility, the threshold values are given by $[0.02, 0.05, 0.06, 0.12, 0.28, 0.40, 0.53, 0.61, 0.73, 0.89, 0.96, 0.99]$. The wide range of threshold values is quite robust to even extreme forms of risk preferences. For instance, for $r_p = 0.99$ ("highly risk averse" according to Holt and Laury (2002)), the range of thresholds is given by $[0.05, 0.07, 0.11, 0.15, 0.18, 0.28.0.34, 0.40, 0.48, 0.60, 0.72, 0.89, 0.98, 0.99]$. For $r_p = -0.99$ ("highly risk loving" according to Holt and Laury (2002)) the thresholds are given by $[0.02, 0.03, 0.06, 0.08, 0.09, 0.19, 0.27, 0.41, 0.63, 0.48, 0.90, 0.72, 0.94, 0.99]$. This suggests that, given r_p is identified, the principals' bonus decisions identify a parameter range of possible values of λ in the deterministic model. In the random utility model, λ is point-identified.

To identify the principals' parameter of risk aversion r_p, subjects in the role of principal completed three multiple price lists (MCL). The MCLs were designed to identify r_p independent of correlation-sensitive preferences, which is taken to assume the form of regret aversion here (Loomes and Sugden, 1982; Bleichrodt et al., 2010; Baillon et al., 2015; Somasundaram and Diecidue, 2017). Given that, at the design stage of the experiment, it seemed possible that correlation sensitivity could play an important role, this feature seemed desirable. I consider the case of regret aversion because its functional form allows for a clear distinction between standard risk preferences and regret aversion, the latter of which captures correlation sensitivity. Regret theory preference can be parameterized as follows (Bleichrodt et al., 2010; Baillon et al., 2015)

$$A \succeq B \iff \sum_{s \in S} q_s (u(x_s^A) - u(x_s^B))^{\beta_p} \tag{B.26}$$

Each choice was between a safe option and a risky option with a high payoff (r_h) and a low payoff (r_l). Multiple switching was possible and is treated as noise within the random utility model. Having a safe option ensures that the correlation structure is clear to subjects, which is important to correctly estimate their correlation sensitivity. This design choice also ensures that a switching point between the safe option and the gamble is independent of a given subject's

Safe	r_h	r_l	prefer safe if $r_p \geq$	Safe	r_h	r_l	prefer safe if $r_p \geq$	Safe	r_h	r_l	prefer safe if $r_p \geq$
350	1880	60	0.97	900	1100	400	-2.28	1000	1650	0	-0.38
450	1880	60	0.80	900	1200	400	-1.07	1000	1650	70	-0.36
550	1880	60	0.65	900	1300	400	-0.41	1000	1650	140	-0.31
650	1880	60	0.51	900	1400	400	0.00	1000	1650	210	-0.23
750	1880	60	0.36	900	1500	400	0.29	1000	1650	280	-0.13
850	1880	60	0.21	900	1600	400	0.50	1000	1650	350	0.00
950	1880	60	0.04	900	1700	400	0.66	1000	1650	420	0.18
1050	1880	60	-0.16	900	1800	400	0.79	1000	1650	490	0.41
1150	1880	60	-0.39	900	1900	400	0.90	1000	1650	560	0.73
1250	1880	60	-0.69	900	2000	400	0.98	1000	1650	630	1.17
1350	1880	60	-1.09	900	2100	400	1.05	1000	1650	700	1.80

Table B.7 Parameters of the MPLs and switching values of r_p.

correlation-sensitivity under the chosen specification. This allows for clean identification of the principals' utility curvature r_p. To see this, observe that a given subject prefers the safe option if and only if

$$0.5[u(S) - u(r_h)]^{\beta_p} + 0.5[u(S) - u(r_l)]^{\beta_p} > 0 \iff$$
$$u(S) - 0.5[u(r_h) + u(r_l)] > 0$$

MCL 1 fixes the 50-50 gamble and increases the safe option, whereas MCL 2 (3) fixes the safe option and the low (high) payoff of the gamble while increasing the size of the high (low) of the gamble. The parameter values are chosen to cover a wide range of possible preferences and can be found in table B.7 along with the switching values for the CRRA utility function.

In addition to the MCLs, principals choose between the 9 choice tasks employed in the principal-agent part of the experiment. These choices can be used to identify the principals' degree of correlation sensitivity β_p, which can help address potential concerns related to correlation-sensitivity in ex-ante preferences. See the robustness exercises in section B.5.4 for more detail. Given that r_p is identified, the choices between the dominant and dominated lotteries from the principal-agent part provide a range of switching values for the parameter of correlation sensitivity β_p. For instance, for a risk neutral principal ($r_p = 0$), the threshold values of β_p are given by $[0.12, 0.25, 0.40, 0.79, 0.91, 1.12, 1.34]$.

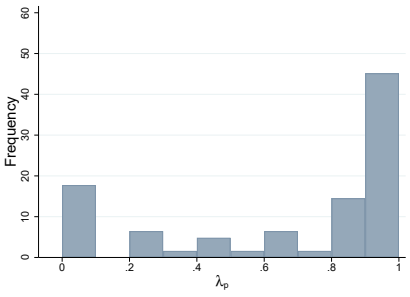
Finally, the noise parameters are identified from how well the specified utility function captures a principal's lottery choices.

Appendix E: Structural estimation- additional results

B.5.1 Principals: Allowing for risk preferences

Figure B.8a present individual-level estimates of λ_p, allowing for preferences to be described by a CRRA utility function. Figure B.8b repeats the regression exercises discussed in the main text for the these estimates. As can be seen, the distribution of degrees of OB remains virtually unchanged,

as do the results from the regression analysis. At the individual level, the estimates of λ_p when imposing $r_p = 0$ or not are almost identical and the two are almost perfectly correlated (Spearman's rho = 0.95). Finally I carry out a likelihood ratio test to test whether imposing $r_p = 0$ reduces the model fit. For this exercise, I focus on the log-likelihood of the bonus decisions. I calculate the log-likelihood for each individual, and the overall log-likelihoods of the restricted and unrestricted model as the sum of the individual likelihood. At the individual level, the unrestricted likelihood of 21 individuals is larger than of the unrestricted model. Since the risk preference parameter is identified by the principals' lottery choices, it is possible that imposing this parameter on the model harms the fit when considering bonus decisions only. The likelihood ratio test does not reject the null that the ratio of the overall likelihoods are different ($\chi^2(62) = 48.182$, $p = 0.901$). This suggests that the unrestricted model, which estimates ρ_p mainly from the principals' lottery choices does not significantly improve, and in some cases even harms, the fit of the bonus decisions.



(a) Individual level estimates of the OB parameter λ_p.

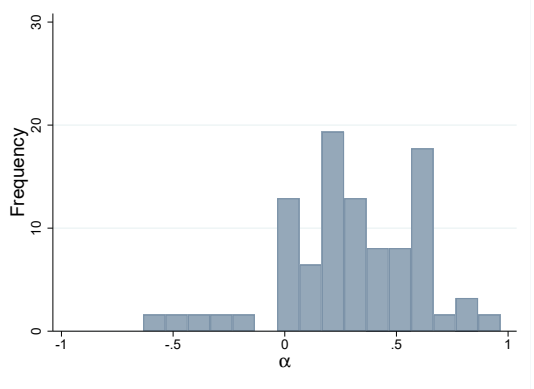
	(1) λ_p	(2) λ_p	(3) λ_p	(4) λ_p	(5) λ_p
CRT score	-0.064**				
	(0.021)				
	[0.015])				
above median seconds		-0.386***			
		(0.083)			
		[0.003])			
"rewarded choices"			0.003		
			(0.022)		
			[0.898]		
"rewarded outcome"				0.067***	
				(0.017)	
				[0.009]	
"rewarded counterfactual"					0.065***
					(0.016)
					[0.009]
Constant	0.869***	0.873***	0.657***	0.264**	0.281**
	(0.070)	(0.037)	(0.164)	(0.128)	(0.119)
Observations	62	62	62	62	62
R-squared	0.123	0.266	0.000	0.217	0.249

(b) OLS regression relating individual level estimates of the OB parameter λ_p to individual characteristics.

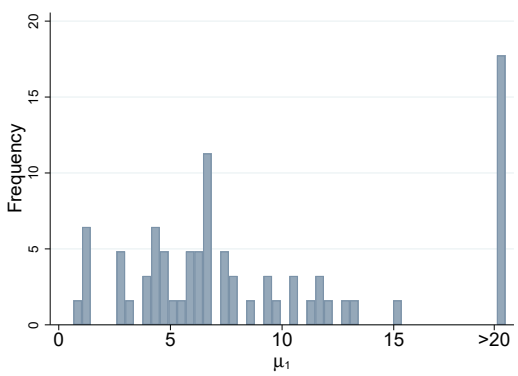
Notes: The estimates allow for risk preferences different from risk neutrality. CRT score ranges from 0-7, with higher values indicating higher cognitive reflection. The "above median seconds" is equal to 1 if a subject's average response time on bonus decisions is above the median. The variables "rewarded ..." indicate to which extent subjects based bonus decisions on the agent's choices, their outcome, or counterfactual comparison between the obtained and the forgone outcome, based on self-reports on a scale from 1-9. Robust standard errors are in parentheses. The sample consists of 62 principals in the reward-after condition. MHT corrected Romano-Wolf step-down p-values in square brackets. *** p<0.01, ** p<0.05, * p<0.1, referring to the MHT corrected p-values (except for the constants). Robust standard errors in parentheses. MHT corrected Romano-Wolf step-down p-values in square brackets. *** p<0.01, ** p<0.05, * p<0.1, referring to the MHT corrected p-values (except for the constants).

Figure B.8 Individual level estimates of the principals' degree of outcome bias.

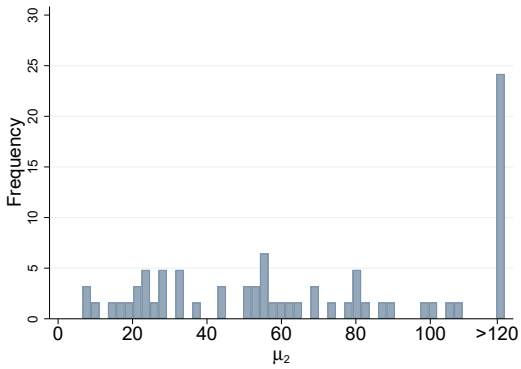
Figure B.9a displays individual-level estimates of principals' risk preference parameter r_p. As can be seen, the estimates center to the right of zero, which reflects the well-documented pattern of a predominance of risk aversion. The modal r_p is 0.30 is consistent with Holt and Laury (2002), who report a modal range of 0.15-0.41 for their low real stakes condition. Figure B.9b and B.9c display individual level estimates of μ_p^1 and μ_p^2. As can be seen, both noise parameters display considerable heterogeneity. Moreover, the estimates of μ_p^2 are much larger than those of μ_p^1. This should not be mistaken to imply that risky choices are captured much better, but the estimation than bonus



(a) Individual level estimates of r_p.



(b) Individual level estimates of μ_p^1.



(c) Individual level estimates of μ_p^2.

decisions. The magnitude in differences in expected utilities is typically much smaller than the magnitude in realized utilities, which makes it difficult to compare the two noise parameters.

Table B.8 displays aggregate level estimates with unrestricted ρ. As can be seen, the estimates are very similar to those for the restricted model with $\rho = 0$. None of the differences are statistically significant ($p > 0.676$, Wald-test). I again perform likelihood ratio tests, using again only the part of the likelihood generated by bonus decisions. The tests suggest that the unrestricted model yields a significantly better fit, for the models in column (1) and (3) ($\chi^2(1) = 46.220$, $\chi^2(1) = 34.117$, and $\chi^2(1) = 39.087$, all $p < 0.001$).

Since the estimates of the degrees of OB do not change significantly when imposing risk neutrality and estimating the unrestricted model does not significantly improve the model fit for the individual-level estimation, I present the results from the restricted model in the main part. This choice also has the advantage of keeping the estimation procedure consistent between principals and agents. If one were to allow for r_p to vary freely on the principals' side, one would have to allow for a non-linear utility function when estimating the agents' anticipated degree of OB. However, the agents' beliefs do not identify their anticipated degree of risk aversion.

	(1) Representative Principal	(2) Full or no OB	(3)
λ	0.615		
	(0.089)		
Fraction $\lambda = 1$		0.705	0.819
		(0.085)	(0.048)
μ^1	4.523	5.534	
	(0.449)	(1.462)	
$\mu^1(\lambda = 0)$			32.913
			(7.996)
$\mu^1(\lambda = 1)$			3.697
			(0.420)
r	0.313	0.312	0.310
	(0.036)	(0.033)	(0.036)
μ^R	24.236	24.248	24.278
	(2.806)	(2.772)	(2.811)
Individuals	62	62	62
Observations	5890	5890	5890
$logL$	-2615.691	-2611.647	-2592.104
BIC	5239.637	5231.549	5196.59
AIC	5235.382	5227.294	5190.209

Table notes: The parameters are estimated from principals' bonus decisions and lottery choices jointly. Column (1) assumes a representative principal. Columns (2) and (3) assume that there is one type with $\lambda = 1$ and another type with $\lambda = 0$, and estimate the fraction of subjects in each category. Column (2) imposes a common noise parameters, whereas column (3) allows for two different noise parameters. Standard errors, in parentheses are obtain from cluster-bootstrap with 4000 repetitions.

Table B.8 Aggregate level estimates of the principals' OB, allowing for risk preferences to differ from risk neutrality.

B.5.2 Principals: Finite mixture models

The estimation of finite mixture models lends further support for the finding that most principals seem to be either fully outcome biased or fully unbiased. A similar estimation exercise on the agents' side thus not yield interesting results and is therefore not reported here.

Table B.9 displays results from the estimation of finite mixture models. As the bootstrap is inconsistent when parameters are at boundary of the parameter space (Andrews, 2000), I provide 95% confidence intervals using the subsample method proposed by Politis and Romano (1994), which is one of the recommendations of Andrews (2000).

The two-type model produces a high OB and a low OB type. λ_1 is estimated at 0.880, and λ_2 at 0.009. While the 95% interval of the former estimate includes 1, that of the latter estimate includes zero. Hence, it cannot be rejected that the first type displays full and the second type no OB. μ^1_1 is estimated at 3.740 whereas μ^1_2 is estimated at 27.122 The fraction of type 1 is 85.5%, whereas that of type 2 is 14.5%. These estimates suggest that the high OB type captures a much larger and somewhat heterogeneous set of principals, whereas the low OB type seems to capture a relatively small set of principals who reward almost exclusively based on the agents' choices and do so quite consistently, as evidenced by the high value of μ^1_2.

The three-type model also produces a high OB type with $\lambda_1 = 1$ and a low OB type with $\lambda_3 = 0$, but also a type with $\lambda_2 = 0.345$. However, this type is imprecisely estimated, with the 95% confidence interval suggesting that the degree of OB could be anywhere between 0.238 and 1.

58.3% are assigned to the high OB type, 30.4% to the imprecisely estimated type 2 and 11.3% to the unbiased type 3.

Finally, the four type model produces one full OB and two low OB types, as well as a type with a degree of OB that is estimated imprecisely. The estimates for type 1 are $\hat{\lambda}_1 = 1$ and $\hat{\mu}_1^1 = 4.806$. Type 2 has $\hat{\lambda}_2 = 0.807$ and $\hat{\mu}_2^1 = 1.457$. The main feature that distinguishes type 2 from 1 is the low value for μ_2^1, which suggest that either heterogeneity within this type or that the posited model does not capture choices well. Consistent with this interpretation the 95% confidence interval for type 2 is given by [0.266, 1]. Type 3 displays low OB ($\hat{\lambda}_3 = 0.227$, 95% CI= $[0.039, 0.437]$) and implements bonus decisions relatively precisely ($\hat{\mu}_3^1 = 8.561$). Type 4 displays virtually no OB ($\hat{\lambda}_3 = 0.005$). The high estimate of the noise parameter $\hat{\mu}_3^1 = 54.988$ suggests that this type follows the rule "reward iff the agent chose the dominant lottery" almost perfectly. 59.7% are allocated to the full OB type 1, while 13.9% are assigned to the low OB type 3, and 11.3% to the unbiased type 4. The remaining 15.1% are assigned to the imprecisely estimated type 2.

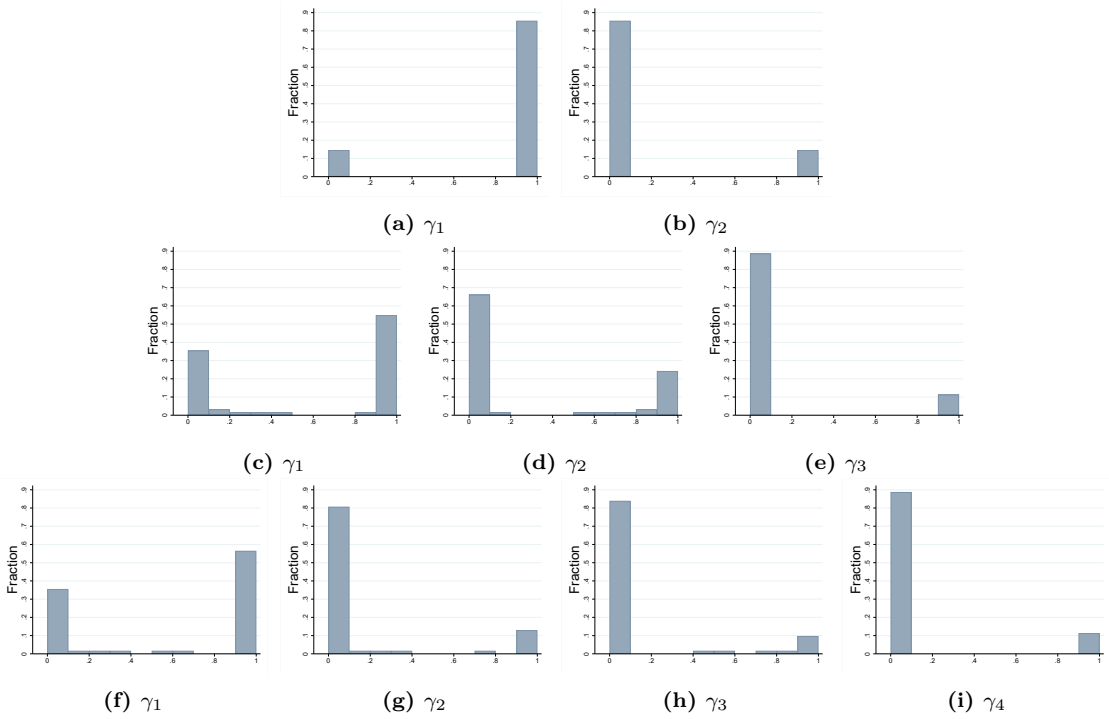
The selection criteria do not agree on which model to select. While the NEC selects the two type model, the ICL selects the four type model. None of the criteria selects the three types model. It can be argued that both the two and four type model share a common feature. None of the models produces a type with an intermediate level of OB. This suggests that most individuals display either low or high OB but not an intermediate level, which confirms the individual-level analysis. These results thus lend support for the claim that individual heterogeneity might be modeled parsimoniously by assuming a fully outcome-biased and fully unbiased type without sacrificing too much realism.

	2 types		3 types		
	type 1	type 2	type 1	type 2	type 3
λ_g	0.880	0.009	1.000	0.345	0.005
	[0.730, 1]	[0, 0.146]	[0.764, 1]	[0.238, 1]	[0.000, 0.036]
μ_g^1	3.740	27.122	4.894	4.080	54.913
	[3.098, 4.784]	[6.478, 88. 375]	[4.443, 5.628]	[0.718, 5.516]	[15.737, 100.547]
γ_g^1	0.855	0.145	0.583	0.304	0.113
	[0.713, 0.935]	[0.065, 0.287]	[0.485, 0.794]	[0.070, 0.374]	[.065, 0.226]
Individuals	62		62		
Observations	3348		3348		
$logL$	-1470.448		-1392.274		
EN	0.001		3.791		
ICL	2961.533		2825.483		
NEC	0.000		0.015		

	4 types			
	type 1	type 2	type 3	type 4
λ_g	1.000	0.807	0.206	0.005
	[0.892, 1]	[0.266, 1]	[0.039, 0.437]	[0.000, 0.018]
μ_g^1	4.806	1.457	8.561	54.988
	[4.358, 5.546]	[0.656, 3.416]	[6.044, 13.228]	[41.282, 108.416]
γ_g^1	0.597	0.151	0.139	0.113
	[0.462, 0.728]	[0.065, 0.254]	[0.058, 0.237]	[0.032, 0.193]
Individuals	62			
Observations	3348			
$logL$	-1374.575			
EN	5.125			
ICL	2803.799			
NEC	0.019			

Table notes: 95% confidence intervals are provided in square brackets. These are obtained using the subsample method proposed by Politis and Romano (1994), with a subsample size of 30 and 4000 samples.

Table B.9 Finite Mixture model of the Principals' degree of outcome bias.



(a) γ_1 **(b)** γ_2 **(c)** γ_1 **(d)** γ_2 **(e)** γ_3 **(f)** γ_1 **(g)** γ_2 **(h)** γ_3 **(i)** γ_4

Figure B.10 The posteriors in the different models.

B.5.3 Agents: Individual-level results

	(1) $\check{\lambda}_p$	(2) $\check{\lambda}_p$	(3) $\check{\lambda}_p$	(4) $\check{\lambda}_p$	(5) $\check{\lambda}_p$
CRT score	-0.000				
	(0.016)				
	[0.999])				
above median seconds		-0.162			
		(0.069)			
		[0.110])			
"made good choices"			-0.007		
			(0.029)		
			[0.970]		
"maximized bonus"				0.029	
				(0.029)	
				[0.669]	
"bad choices to max bonus"					0.023**
					(0.011)
					[0.015]
Constant	0.768***	0.848***	0.822***	0.530**	0.674***
	(0.071)	(0.039)	(0.241)	(0.236)	(0.064)
Observations	67	67	67	67	67
R-squared	0.000	0.079	0.001	0.015	0.055

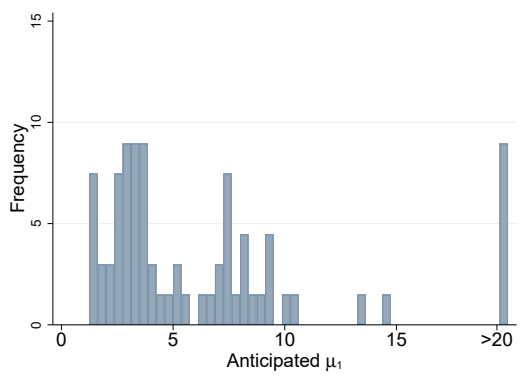
Table notes: CRT score ranges from 0-7, with higher values indicating higher cognitive reflection. The "above median seconds" is equal to 1 if a subject's average response time on bonus decisions is above the median. The variables "rewarded ..." indicate to which extent subjects based bonus decisions on the agent's choices, their outcome, or counterfactual comparison between the obtained and the forgone outcome, based on self-reports on a scale from 1-9. Robust standard errors are in parentheses. The sample consists of 67 agents in the reward-after condition. MHT corrected Romano-Wolf step-down p-values in square brackets. *** p<0.01, ** p<0.05, * p<0.1, referring to the MHT corrected p-values (except for the constants). Robust standard errors in parentheses. MHT corrected Romano-Wolf step-down p-values in square brackets. *** p<0.01, ** p<0.05, * p<0.1, referring to the MHT corrected p-values (except for the constants).

Table B.10 OLS regression relating individual level estimates of the anticipated degree of OB $\check{\lambda}_p$ to individual characteristics.

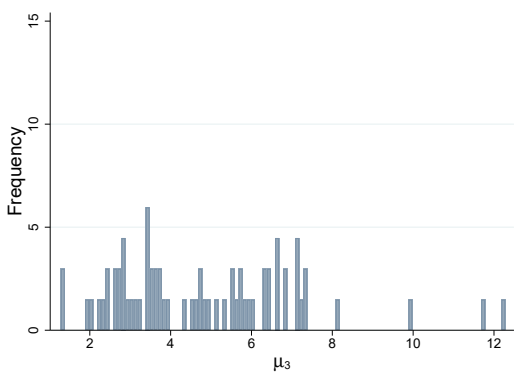
As for the principals' degree of OB, I explore correlations of the estimated anticipated degree of OB with individual characteristics. Overall, individual characteristics do not explain variation in the estimated $\check{\lambda}_p$ well. I explore correlations with participants' CRT score, response time, and survey responses indicating to which extent they tried to "make a good choice", "chose an option that maximized their chances of obtaining a bonus", and "sometimes chose a bad option if that maximized the chances of obtaining a bonus".[17] Results are reported in table B.10, columns (1)-(5). Before correcting for MHT, only response time and reporting to sometimes having made a bad choice when this maximized the probability of obtaining a bonus are significant at the 5% level. After controlling for MHT, none of the estimated coefficients retains statistical significance. Moreover, the r-squares are significantly lower than for the regression exercises conducted on the principals' OB, within none exceeding 0.082.

The lack of significant associations between anticipated OB and measures of individual characteristics could be due to two reasons. First, heterogeneity is decisively less pronounced in the estimates of the anticipated degree of OB. As such, it might not be surprising that the heterogeneity in OB correlates with measures of individual characteristics. Second, it is possible that the measured dimensions of individual characteristics are simply more relevant for OB than antic-

[17] Subjects' responses were recorded on a 9-point Likert scale, where "1" indicated no agreement and "9" full agreement.



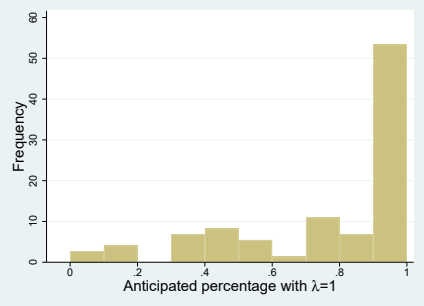
(a) Individual level estimates of $\check{\mu}_1$.



(b) Individual level estimates of μ_3.

ipated OB.[18] The finding that reported strategies for choices largely do not predict the anticipate outcome bias well could result from the fact that choices are largely inconsistent with beliefs.

Results when imposing that agents imagine that principals are either fully outcome biased or fully unbiased can be found in figure B.11a and table B.11b. As can be seen results are, overall fairly similar. However, the association between 'sometimes chose a bad option if that maximized the chances of obtaining a bonus" and anticipated outcome bias becomes insignificant, which suggests that the correlation is not very robust.



(a) Individual level estimates of the agents' beliefs over $\check{\lambda}$.

	(1) $\check{\lambda}_p$	(2) $\check{\lambda}_p$	(3) $\check{\lambda}_p$	(4) $\check{\lambda}_p$	(5)
CRT score	0.006				
	(0.015)				
	[0.904])				
above median seconds		-0.157			
		(0.060)			
		[0.084])			
"made good choices"			-0.009		
			(0.029)		
			[0.904]		
"maximized bonus"				0.018	
				(0.025)	
				[0.850]	
"bad choices to max bonus"					0.017
					(0.009)
					[0.292]
Constant	0.770***	0.869***	0.865***	0.645***	0.721***
	(0.066)	(0.034)	(0.246)	(0.208)	(0.057)
Observations	67	67	67	67	67
R-squared	0.002	0.096	0.002	0.007	0.040

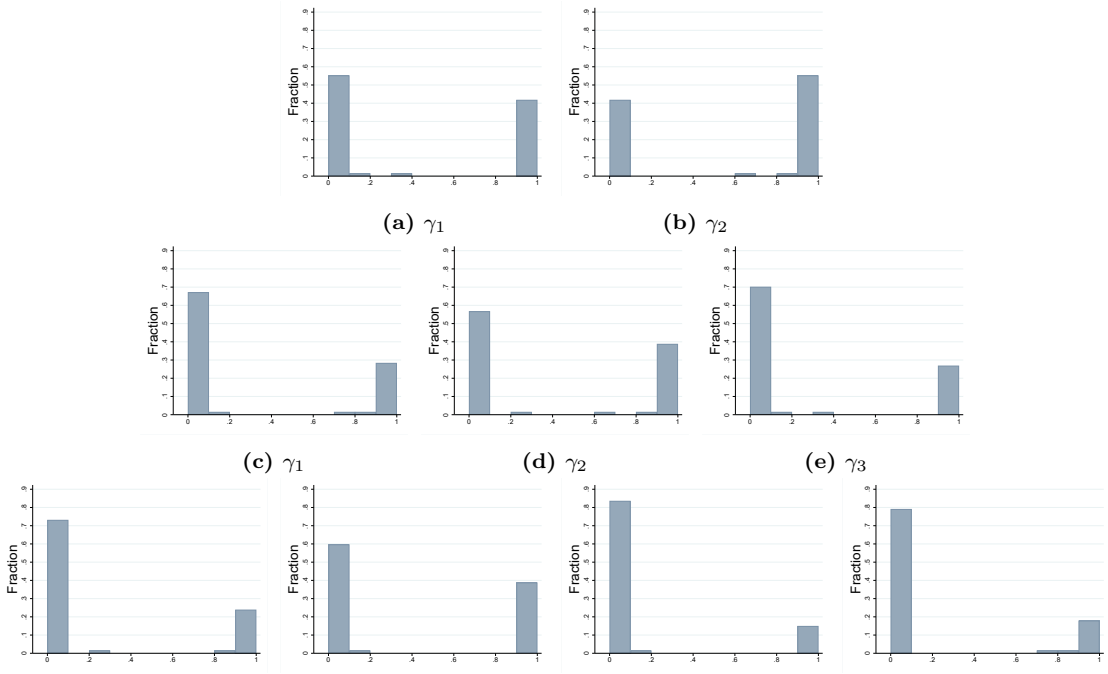
(b) Robust standard errors in parentheses. MHT corrected Romano-Wolf step-down p-values in square brackets. *** p<0.01, ** p<0.05, * p<0.1, referring to the MHT corrected p-values (except for the constants).

Figure B.11 Individual level estimates of the agents' anticipated degree of outcome bias.

B.5.4 Robustness and extension of the model

I employ the same estimation framework based on a random utility approach and maximum likelihood as for the baseline model discussed in section 4.6. Hence, the probability with which a principal awards the bonus is still given by

[18] One might argue that the CRT score or response time might better correlate with the deviation of $\check{\lambda}_p$ from the principals' true degree of OB. However, regression analysis (not reported here) suggests that this is not the case.



(a) γ_1 **(b)** γ_2 **(c)** γ_1 **(d)** γ_2 **(e)** γ_3 **(f)** γ_1 **(g)** γ_2 **(h)** γ_3 **(i)** γ_4

Figure B.13 The posteriors in the different models.

$$pr(\lambda_p, r_p, \mu_p^1, \theta_k^{corr}, s) = \frac{1}{1 + exp(-\mu_p^1 z(\theta_k^{corr}, s))}, \tag{B.27}$$

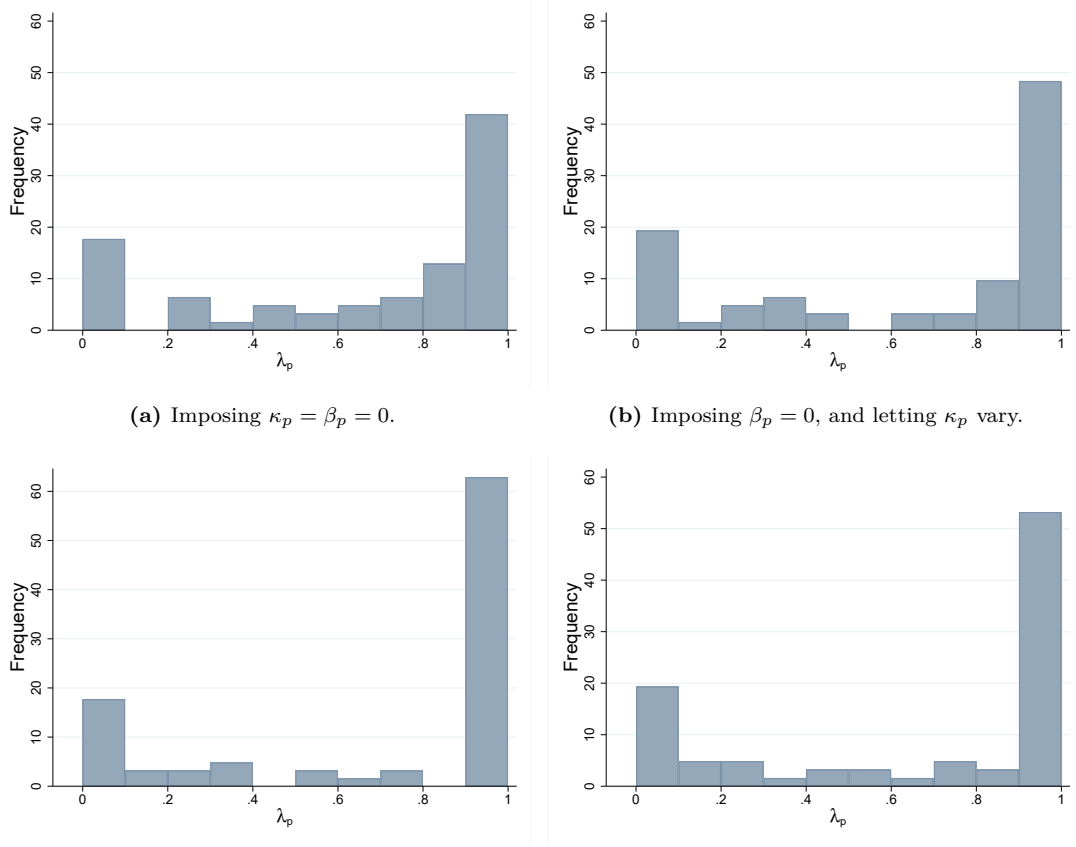
but I modify $z(\theta_k^{corr}, s)$ to allow for bonus decisions to depend on the obtained outcome and for preferences to be correlation-sensitive.

$$z(\theta_k^{corr}, s) = \lambda_p \phi(x_s^{\theta_k^{corr}}, x_s^{-\theta_k^{corr}}) + (1 - \lambda_p)\Phi(\theta_k^{corr}, -\theta_k^{corr}) + \kappa u(x_s^{\theta_k^{corr}}), \tag{B.28}$$

where κ captures to which extent bonus decisions are influenced by the obtained outcome. This parameter could capture, for instance, outcome based reciprocity, or joy from obtaining a higher outcome that induces principals to act more kindly towards the agent. The function $\phi(\cdot)$ captures correlation-sensitivity in preferences. Following research on regret theory (Loomes and Sugden, 1982; Bleichrodt et al., 2010; Baillon et al., 2015; Somasundaram and Diecidue, 2017) I impose $\phi(x_s^{\theta_k^{corr}}, x_s^{-\theta_k^{corr}}) = (u(x_s^A) - u(x_s^B))^{\beta_p}$, where β_p measures and individual's degree of correlation sensitivity and $u(\cdot)$ is defined as before. In regret theory $\beta_p > 1$. $\phi(x_s^{\theta_k^{corr}}, x_s^{-\theta_k^{corr}})$ can be thought of as the net benefit of obtaining a payoff $x_s^{\theta_k^{corr}}$ instead of $x_s^{-\theta_k^{corr}}$. I define $\Phi(\theta_k^{corr}, -\theta_k^{corr}) = \sum_{s \in S} p_s \phi(x_s^{\theta_k^{corr}}, x_s^{-\theta_k^{corr}})$, which can be though of as the ex-ante benefit of choosing lottery θ instead of lottery $-\theta$. Note that this model nests the baseline model for $\beta_p = 0$ and $\kappa_p = 0$.

I repeat the estimation exercises reported in section 4.6 imposing $\kappa_p = 0$ or $\beta_p = 0$ or letting both vary freely. Figure B.14 shows the distribution of the estimates. For ease of comparison, I also include the baseline model that imposes $\kappa_p = \beta_p = 0$. As can be seen, the distributions of the estimated parameters are very similar across the different specifications and maintain the two spikes near 0 and 1. Table B.15 shows coefficients obtained when conducting regression analysis

equivalent to that discussed in section 4.6 for the different model specifications. As can be seen, the coefficients are nearly identical for the different specifications. Figure B.16 provides a similar exercise for the individual-level estimates for the agents.



(a) Imposing $\kappa_p = \beta_p = 0$.

(b) Imposing $\beta_p = 0$, and letting κ_p vary.

(c) Imposing $\kappa_p = 0$, and letting β_p vary.

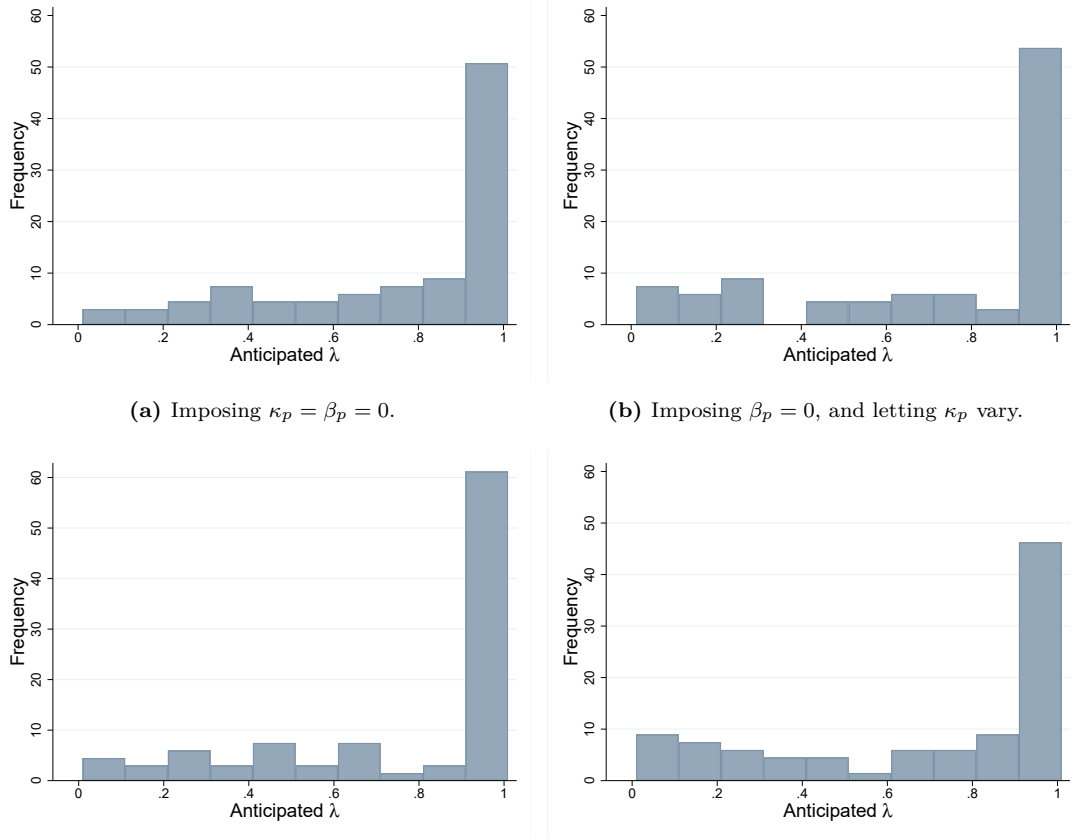
(d) Letting both κ_p and β_p vary.

Figure B.14 Individual level estimates of the principals' degree of outcome bias for different model specifications.

	$\kappa_p = \beta_p = 0$	$\beta_p = 0$	$\kappa_p = 0$	full
CRT score	-0.065**	-0.066**	-0.062*	-0.060*
	(0.021)	(0.024)	(0.024)	(0.024)
	[0.012]	[0.032]	[0.060]	[0.079]
above median seconds	-0.386***	-0.401***	-0.400***	-0.377***
	(0.083)	(0.089)	(0.089)	(0.093)
	[0.003])	[0.004]	[0.005]	[0.012]
"rewarded choices"	0.004	0.001	-0.005	-0.009
	(0.022)	(0.024)	(0.026)	(0.027)
	[0.878]	[0.967]	[0.848]	[0.735]
"rewarded outcome"	0.066***	0.064**	0.069***	0.067**
	(0.017)	(0.019)	(0.018)	(0.018)
	[0.009]	[0.018]	[0.013]	[0.019]
"rewarded counterfactual"	0.064***	0.065***	0.071***	0.068**
	(0.016)	(0.017)	(0.017)	(0.017)
	[0.009]	[0.014]	[0.010]	[0.012]
Observations	62	62	62	62

Notes: Robust standard errors in parentheses. MHT corrected Romano-Wolf step-down p-values in square brackets. *** p<0.01, ** p<0.05, * p<0.1, referring to the MHT corrected p-values (except for the constants).

Figure B.15 Coefficients from OLS regressions that regress the estimated λ_p from different model specifications on individual characteristics.



(a) Imposing $\kappa_p = \beta_p = 0$.

(b) Imposing $\beta_p = 0$, and letting κ_p vary.

(c) Imposing $\kappa_p = 0$, and letting β_p vary.

(d) Letting both κ_p and β_p vary.

Figure B.16 Individual level estimates of the agents' anticipated degree of outcome bias for different model specifications.

I further conduct aggregate level estimation with the additional specification. As κ and β might vary across outcome biased and unbiased principals, I report here only results assuming a representative principal. For convenience, I also report the estimates of the baseline model that imposes $\kappa = \beta = 0$. As can be seen in Tables B.11 and B.12, the estimates of the principals' degree of OB and the agents' anticipated degree of OB are decreased somewhat when allowing for correlation-sensitive risk preferences and bonus decisions that depend on the principal's obtained payoff separately from the ex-post payoff comparison. None of the estimates of the degree of OB of the representative principals differ significantly from the baseline model ($p > 0.271$, Wald test). For the agents, the baseline estimate of $\check{\lambda} = 0.788$ does not differ significantly from the estimates when allowing either $\check{\kappa}$ or $\check{\beta}$ to vary freely ($p = 0.756$ and $p = 0.344$ respectively, Wald test) but it does differ at $p = 0.051$ from the estimate of 0.610 when allowing both parameters to vary freely. In summary, the aggregate estimates of OB and anticipated OB are somewhat reduced, but the qualitative conclusions drawn from the baseline model do not change.

	(1)	(2)	(3)	(4)
	Representative Principal			
λ	0.589	0.502	0.555	0.456
	(0.085)	(0.096)	(0.080)	(0.085)
μ^1	4.524	4.246	4.488	4.220
	(0.417)	(0.499)	(0.428)	(0.497)
κ		0.438		0.445
		(0.075)		(0.078)
β			0.890	0.835
			(0.057)	(0.065)
Individuals	62	62	62	62
Observations	3348	3348	3348	3348
$logL$	-1648.012	-1464.591	-1646.444	-1461.633
BIC	3304.278	2934.559	3298.265	2930.435
AIC	3300.023	2935.182	3298.888	2931.266

Table notes: Column (1) imposes $\kappa = \beta = 0$. Column (2) imposes $\beta = 0$, and Column (3) imposes $\kappa = 0$, whereas column (4) imposes no restrictions. Standard errors, in parentheses are obtain from cluster-bootstrap with 4000 repetitions.

Table B.11 Aggregate level estimates of the principals' OB, assuming a representative principal.

	(1)	(2)	(3)	(4)
	Representative Agent			
$\check{\lambda}$	0.788	0.755	0.699	0.610
	(0.064)	(0.083)	(0.068)	(0.065)
$\check{\mu}^1$	3.872	3.306	3.733	3.205
	(0.323)	(0.319)	(0.312)	(0.305)
μ^2	3.439	3.892	3.449	3.911
	(0.167)	(0.189)	(0.170)	(0.189)
$\check{\kappa}$		0.524		0.561
		(0.094)		(0.103)
$\check{\beta}$			0.802	0.687
			(0.065)	(0.059)
Individuals	67	67	67	67
Observations	3618	3618	3618	3618
$logL$	-1254.479	-861.9919	-1250.494	-847.9459
BIC	2517.367	1731.288	2508.292	1705.022
AIC	2512.958	1731.983	2508.988	1705.891

Table notes: Column (1) imposes $\kappa = \beta = 0$. Column (2) imposes $\beta = 0$, and Column (3) imposes $\kappa = 0$, whereas column (4) imposes no restrictions. Standard errors, in parentheses are obtain from cluster-bootstrap with 4000 repetitions.

Table B.12 Aggregate level estimates of the agents' anticipated OB, assuming that the Representative agents imagines a representative principal.

Appendix F: Preanalysis plan

The following replicates the preanalysis submitted to the AER registry plan word for word.

B.6.1 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), that 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 eliminate incentives to choose optimal actions? 2) Do agents anticipate the OB of principals correctly and 3) do they adjust their choices accordingly, i.e. can outcome bias induce more choices of sub-optimal actions and thus decrease welfare?

B.6.2 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. Treatments are assigned at the session level. 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 54 bonus decisions in the reward-after treatment and a

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

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

Table notes: Rows 2 and 3 display the payoffs of option G (FOSD) and B in the different states of the world.

Table B.13 The first row presents different possible states and their probability of occurring.

total of 18 bonus decisions in the reward-before treatment. See table B.13 and B.14 for the lottery tasks employed. In both treatments, the bonus amounts to 10 Euros. Whenever the bonus is not allocated to the agent matched to 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 lottery 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.

Each agent makes 2 choices for each of the 9 lottery pairs. In addition, beliefs are elicited. Agents make a first choice for each of the 9 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.

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 agent and the bonus decision of the principal is implemented.

In a addition to the above, principals also make choices between the lotteries used in part I of the experiment for themselves. In addition, 3 multiple choice lists are used to elicit risk preferences.

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).

All subjects will further answer a questionnaire (see section B.6.2.1).

The experiment will be conducted at the lab at Toulouse School of Economics, in April 2023. The target number for participants is 300, with 140 participants in the reward-before and 160 participants in the reward-after treatment, and an equal number of principals and agents. The target is to have at least 8 participants in each session.

Participants will receive a show-up fee of 5 Euros and expected additional earnings of around 10 Euros for an expected duration of 60-75 minutes.

B.6.2.1 Questionnaire items

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

- An extended version of the cognitive reflection test (Frederick, 2005; Toplak et al., 2014)

lottery pair	H	M	L	ϵ
1	1953	1031	109	45
2	1953	1031	109	110
3	1403	688	103	359
4	1403	688	103	523
5 (corr 2 only)	1480	750	50	699

Table B.14 Parameter values for the different lotteries. All payoffs are in cents.

- Willingness to take risk (Dohmen et al., 2011)
- Standard demographics: These include age, gender, field of study, nationality, level of education, and household income.

Moreover, subjects in the role of principal are asked to which 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 which 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.

B.6.3 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 $V(\theta) = \sum_{s \in S} p_s v(x_s^\theta)$.[19]

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

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

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 $lp \in \{1, 2, 3, 4\}$, from a principal $j \in J$, after lottery choice $\theta \in \{G, B\}$ under correlation

[19] $v()$ 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.

structure $corr \in \{1, 2\}$.[20] 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 implications of the model are the following. First, $I_{lp,j}(G^{corr})$ is weakly decreasing in λ_j. Second, even if $\lambda_j = 1$, agents have positive incentives to choose the dominant lottery under correlation 1, i.e. $I_{lp,j}(G^1) > 0$ for all $lp \in \{1, 2, 3, 4\}$. However, under correlation 2, there exists $\overline{\lambda}_{lp}$ such that $I_{lp,j}(G^2) < 0$ for all $\lambda_j \in (\overline{\lambda}_{lp}, 1]$. Moreover, the choice tasks are chosen such that $\overline{\lambda}_1 < \overline{\lambda}_2 < \overline{\lambda}_3 < \overline{\lambda}_4$. It further holds true that $I_{1,j}(G^{corr}) \leq I_{2,j}(G^{corr}) \leq I_{3,j}(G^{corr}) \leq I_{4,j}(G^{corr})$, for $corr \in \{1, 2\}$.[21] In the experiment, agents are randomly matched to principals. Their incentives to choose lottery G are therefore given by $I_{lp}(G^{corr}) = \frac{1}{|J|} \sum_{j \in J} I_{lp,j}(G^{corr})$. The different lottery pairs allow to gauge the strength of correlation sensitivity, both at the individual and the aggregate level (see section B.6.5 for further discussion).

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.

B.6.4 Main Hypotheses

All null hypotheses in this section will be tested at the 5% significance level. For within subject tests, I will use the Wilcoxon signed-rank test. For between subject tests, I will use the Wilcoxon rank-sum test.

B.6.4.1 Variable construction

Denote $P_{lp,j}(\theta^{corr}, T)$ the probability of receiving the reward for lottery pair $lp \in \{1, ..., 4\}$ 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 probabilities. Hence $P_{lp,j}(C^{corr}, before) \in \{0, 1/3, 2/3, 1\}$.

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. Tests will be carried out for the different lottery pairs separately.

B.6.4.2 Preliminary: Outcome bias in bonus decisions

Hypothesis 8. ***Outcome bias in bonus decisions:*** *Principals are more likely to award the bonus if*

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

[21] This assumes that v is a v.N.M. utility function with reasonable parameters of risk aversion.

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.

$$Bonus_{i,t} = \beta_0 + \beta_1 preferred_{i,t} + \beta_2 obtained\ payoff_{i,t} + \beta_3 \{obtained > forgone\}_{i,t} + \epsilon_{i,t}, \tag{B.29}$$

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

- For hypothesis 8a) I test the null hypothesis that $\beta_1 = 0$, against the alternative hypothesis that $\beta_1 > 0$, for both treatments
- For hypothesis 8b) 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").[22]

B.6.4.3 OB and incentives to choose lottery G

In hypothesis 9, I collect the hypotheses on how OB affects the agents' incentives to choose the dominant lottery.

Hypothesis 9. ***Outcome bias and incentives-aggregate level:***

a) $I(G^1, before) > 0$ *and* $I(G^2, before) > 0$*. Moreover, no significant difference arises between* $I(G^1, before)$ *and* $I(G^2, before)$*.*

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 9a) I test the null hypotheses that $P(G^{corr}, before) = P(B^{corr}, before)$, for $corr \in \{1, 2\}$, against the alternative hypothesis that $P(G^{corr}, before) > P(B^{corr}, before)$. I also test the null hypothesis that $I(G^1, before) = I(G^2, before)$, against the alternative $I(G^1, before) \neq I(G^2, before)$. Wilcoxon signed-rank tests will be used.

[22] 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.

- For hypothesis 9b), 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 9c), 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.

B.6.4.4 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 $treatment \in \{before, after\}$. I average the choice frequency for both choices agents make for each choice task.

Hypothesis 10. ***The agents' choices***

a) In the reward-before treatment, agents choose the dominant lottery at a high frequency, under both correlation structures. Moreover, no significant difference arises between $F(G^1, before)$ *and* $F(G^2, before)$.

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

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

I will test the following null hypotheses.

- For hypothesis 10a) I test the null hypotheses that $F(G^1, before) = F(G^2, before)$ against the alternative that $F(G^1, before) \neq F(G^2, before)$, using a Wilcoxon signed-rank test. The hypothesis is not expected to be rejected.

- For hypothesis 10b) I test the null hypotheses that $F(G^{corr}, before) = F(G^{corr}, after)$, against the alternative that $F(G^{corr}, before) > F(G^{corr}, after)$, for $corr \in \{1, 2\}$, using a Wilcoxon rank-sum test.

- For hypothesis 10c) I test the null hypotheses that $F(G^1, before) - F(G^2, before) = F(G^1, after) - F(G^2, after)$, against the alternative that $F(G^1, before) - F(G^2, before) < F(G^1, after) - F(G^2, after)$ using a Wilcoxon rank-sum test.

B.6.5 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. If the principals' preferences are found to be strongly correlation-sensitive (which is not expected), hypotheses 9 and 10 should not be expected to hold true.

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.

Further, the different lottery pairs allow to gauge the usefulness of the model. The model predicts that OB should lead to greater changes in incentives and lottery choices for lottery pairs with lower indices. Therefore, hypotheses 8b), 9 b) and c), and 10 b) and c) (given that agents form accurate beliefs) are most likely to hold true for $lp = 1$, are somewhat less likely for $lp = 2$, and are more unlikely for $lp = 3$ and even less likely for $lp = 4$, since higher and higher levels of outcome bias are required.

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 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.

Appendix G: Details of the Experiment

B.7.1 Screenshots

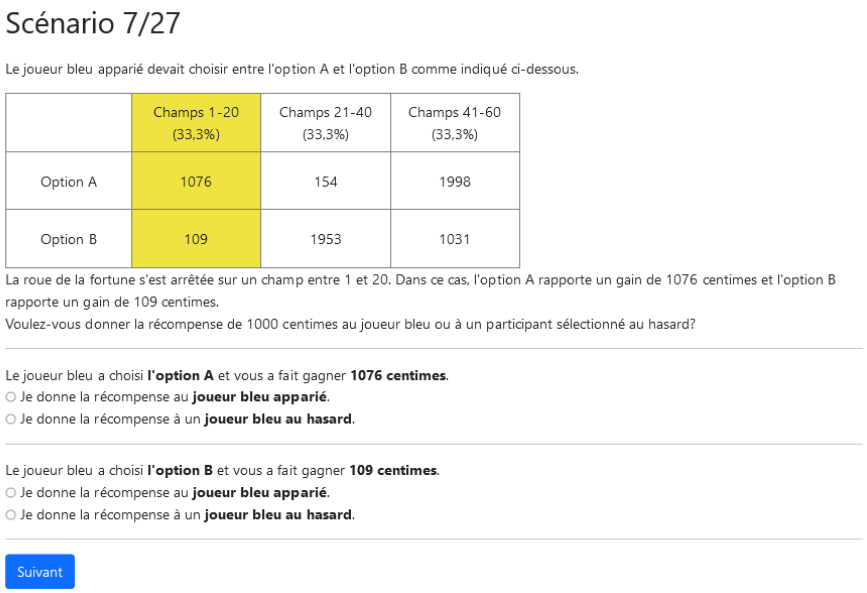


Figure B.17 Example screen shot of bonus decisions in the reward-after treatment.

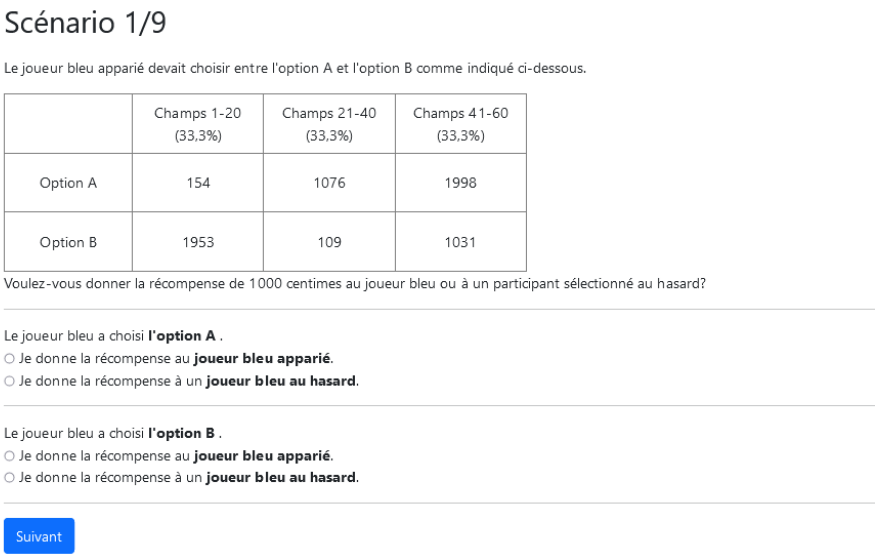


Figure B.18 Example screen shot of bonus decisions in the reward-before treatment.

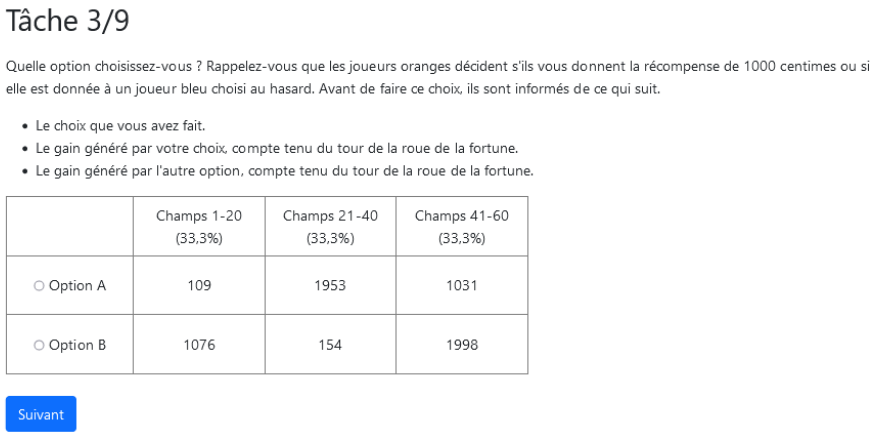


Figure B.19 Example screen shot of the agents' choice screen in the reward-after treatment.

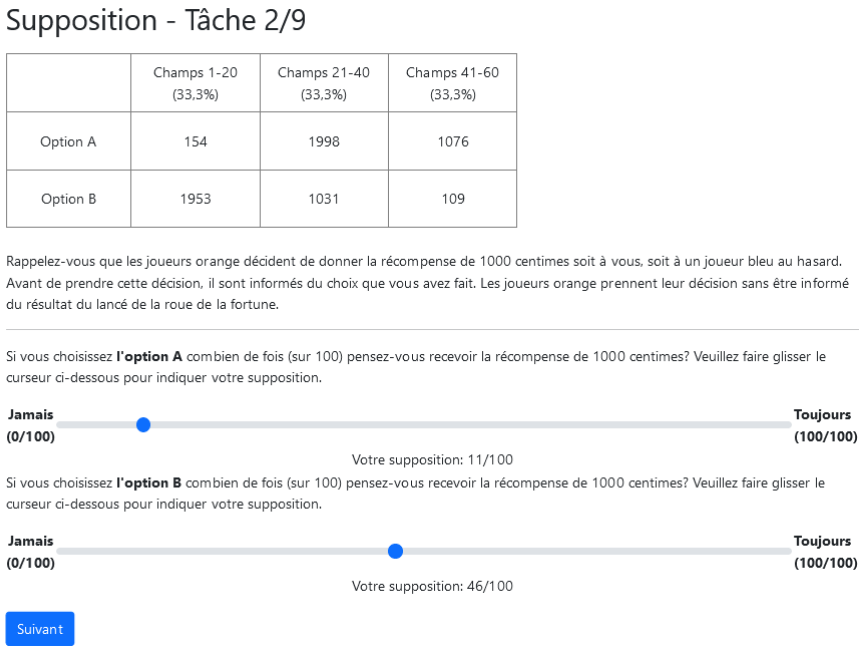


Figure B.20 Example screen shot of the agents' choice screen in the reward-before treatment.

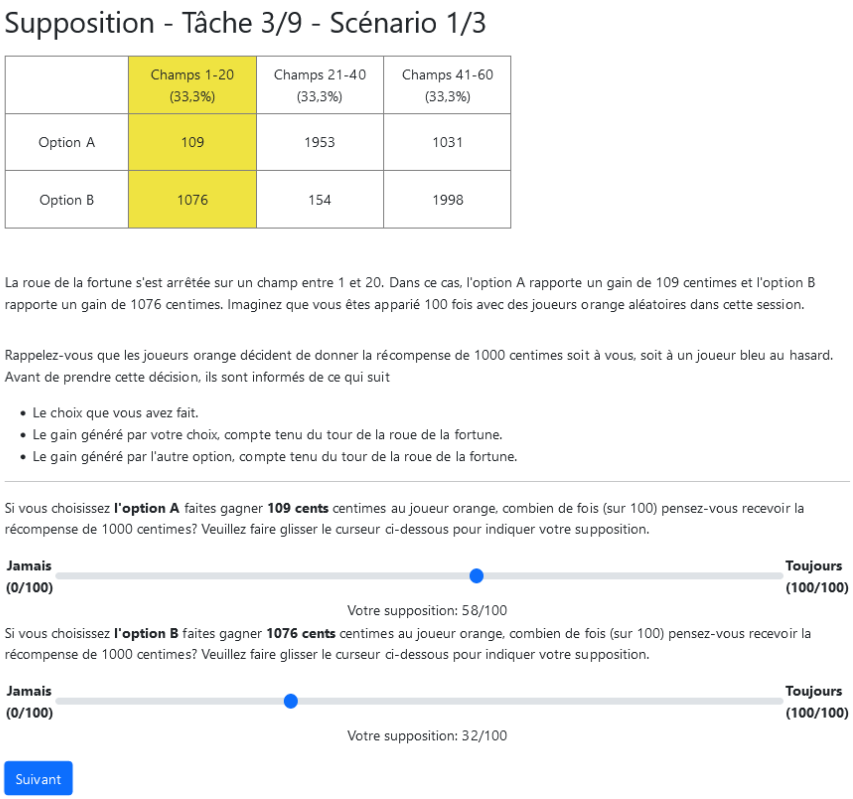


Figure B.21 Example screen shot of the agents' choice screen for the first and last round choices in the reward-after treatment.

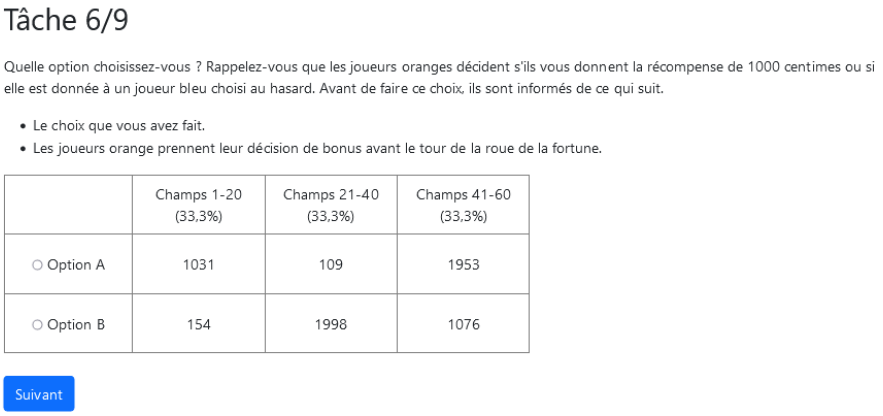


Figure B.22 Example screen shot of the agents' choice screen for the first and last round choices in the reward-before treatment.